SCALING LAWS FOR SIMULATION FIDELITY: SCALE ALONE IS NOT ENOUGH

Anonymous authors
Paper under double-blind review

ABSTRACT

Both humans and AI can learn new skills by practicing with simulated human partners. Current LLM simulations can already provide meaningful learning signals, but they remain significantly less effective than human partners. Although LLM simulations look plausible on the surface, they often fail to reproduce human behavior patterns that are necessary for realistic and effective practice. We call this a gap in *simulation fidelity*. First, we introduce and validate distributional fidelity metrics, which we use to quantify fidelity gaps in two representative simulation tasks. We show that domain adaptation fails to close these gaps. This motivates our investigation of scaling laws for simulation fidelity. Web pre-training data is cheaper and more scalable than in-domain data, so we systematically study how simulation fidelity scales with pre-training compute on webtext. We discover that fidelity scales log-linearly with pre-training at smaller scales, but fidelity quickly asymptotes. Relative to other skill domains like creative writing and open-ended question answering, the rate of simulation performance improvement is much slower. Furthermore, domain adaptation does not alter the rate at which pre-training compute benefits model fidelity. In summary, we conclude that, even at scale, webtext data is an insufficient grounds for building more faithful LLM simulators. We recommend greater investment in larger-scale in-domain datasets than what are currently available.

1 INTRODUCTION

LLM-simulated human partners have become increasingly useful teaching tools (Nguyen & Cao, 2026; Müller et al., 2026). Educators can practice teaching simulated students (Pan et al., 2025), therapists practice counseling (Louie et al., 2026), clinicians practice patient encounters (Yu et al., 2025), and everyday users practice soft skills (Yang et al., 2024). Simulated partners can also provide AI agents with RL training practice to become better teachers (Scarlatos et al., 2026) and collaborators (Sun et al., 2025; Wu et al., 2025). In these settings, LLM partners are already providing meaningful learning signals, but they remain significantly less effective than real human partners.

Simulations are less effective if they lack *fidelity*, or fail to reproduce the human interaction behaviors that are necessary to practice the target skill. For example, LLM partners are often overly cooperative (Chopra et al., 2026), and they fail to provide the necessary resistance for students to learn skills like therapy (Tian et al., 2026), tutoring (Martynova et al., 2025), and conflict resolution (Kwon et al., 2026). LLM simulations also lack cognitive plausibility (Zhou et al., 2026b; Scarlatos et al., 2025) and psychological consistency (Prabhakar et al., 2026; Wang et al., 2025), making simulations unrealistic. Low-fidelity simulations are less effective at teaching professionals (Donovan et al., 2025; Luebbert et al., 2023), and less effective at training AI assistants (Suh et al., 2026).

Progress towards more faithful LLM simulators is limited first by the lack of reliable evaluation metrics (Scarlatos et al., 2026; Liu et al., 2026). As our first contribution in this paper, **we contribute and validate distributional metrics** for simulation fidelity (§2), which we use **to quantify fidelity gaps** in two representative simulation tasks: *user simulation* and *patient simulation*. Specifically, we show that frontier models are unfaithful simulators (§3) that over-act and lack resistance. For this reason, even small 2.5B base models can outperform frontier models at patient simulation. But base models are also significantly worse than human partners. Even domain adaptation with up to 15M tokens of available data fails to significantly close these gaps.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Webtext pre-training data is a cheaper and more abundant alternative to in-domain training data. **Our principal research question is whether pre-training on webtext provides consistent and reliable benefits to simulation fidelity at scale,** such that it can replace the need for more in-domain training data (§4). We formalize this investigation by fitting carefully-controlled compute scaling laws. Prior work shows strong evidence of log-linear scaling laws in more constrained MCQ-style simulation settings where log-likelihood-based scaling is straightforward (Ziems et al., 2026). But these results do not immediately transfer to the open-ended domain of partner simulations. For example, prior work found that larger models are not always more faithful user simulators (Suh et al., 2026).

We find that simulation fidelity scales with pre-training compute at small scales ($10^{18} - 10^{20}$ FLOPs), with tight log-linear fits in each domain respectively ($r^2 = 0.98$, $r^2 = 0.88$). This lets us express the value of in-domain data in units of comparable compute benefit. The gain from domain adaptation of small models with 15M tokens corresponds to as little as $1.5\times$ the pre-training compute for user simulation. Pre-training and in-domain adaptation are complementary levers, but especially where in-domain data is less abundant, pre-training can offer a scalable complement.

However, the benefits of pre-training alone has diminishing returns. **We observe asymptotic limits to simulation fidelity that are significantly worse than human level**. To study this further, we establish the correlation between fidelity and loss, and show that with respect to loss, the rate of improvement with scale is relatively much slower than what we see in other open-ended language generation domains, like creative writing and open-ended question answering (§5). We quantify these disparities in *relative scaling laws* (Held et al., 2025) and show that scaling disparities are not mitigated by domain adaptation. We conclude that webtext is an insufficient ground for building more faithful LLM simulators. We recommend greater investment in the development and large-scale curation of in-domain datasets for training LLMs as partner simulators.

2 Simulation Tasks and Fidelity Metrics

To be effective training systems, LLM partner simulations need to be faithful models of human behavior (Donovan et al., 2025; Luebbert et al., 2023; Suh et al., 2026). Progress towards faithful simulators has been limited by the lack of automated evaluation metrics (Scarlatos et al., 2026; Liu et al., 2026). For this reason, we introduce and validate new fidelity metrics that align with the education literature. Here, fidelity is the degree of accuracy with which a simulation represents the given frame of reality that is most conducive to educational effectiveness (Tun et al., 2015). The relevant frame is domain-specific, so we consider two representative tasks domains (§2.1). For each task domain, we operationalize fidelity by decomposing the relevant frame into measurable behavioral features (§2.2). We then project human and model responses into this shared feature space and compute their distributional similarity, which serves as our fidelity metric (§2.3).

2.1 Simulation Tasks

We formulate partner simulation as conditional next-turn generation from an interaction log with alternating turns $(a_1, p_1, ..., a_N, p_N)$ between an agent a and their interaction partner p. Given a real conversation history h and optional social context c, the task of the LLM-simulator is to generate the next partner turn $\hat{y} \sim P_m(y \mid h, c)$. We evaluate each turn independently from real histories rather than feeding simulated responses back into the conversation to avoid error accumulation.

There are two major uses for LLM-based partner simulations. They can be used to train and evaluate AI, or they can be used to train real human practitioners. For each use case, we select a representative task: *user simulation* for AI training and *therapy patient simulation* for human training. We ground both tasks in naturalistic human interaction data.

For **user simulation**, we use the PRISM Alignment Dataset (Kirk et al., 2024), which contains real interactions between AI assistants a and human users p from a wide variety of demographic backgrounds. For each pair (a, p), we randomly sample from the full interaction log a single-turn history $h = (p_i, a_i)$ and its ground truth partner response p_{i+1}. In total, we evaluate on 456 samples with 5 generated responses per sample.

For **patient simulation**, we use the Alexander Street Press (2013) *Counseling and Psychotherapy Transcripts*, which includes 1,841 full-text, anonymized transcripts from real dialogues between

therapists a and patients p across a wide variety of clinical settings and therapeutic methods. For each pair (a, p), we sample a six turn history $h = (p_i, a_i, ..., p_{i+2}, a_{i+2})$ and its ground truth patient response p_{i+3}, ensuring that the target utterance contains at least 60 characters. In total, we evaluate on 1,826 samples. For social context c, we include the therapist's therapeutic approach, and the patient's age range, gender, marital status, and presenting problem.

2.2 Relevant Behavioral Features

A good fidelity metric should account for the degree to which a simulation represents the behavioral feature variables most *relevant* to effective and realistic practice. We identify and operationalize the most relevant features for each task, as shown in Table 1. For each feature, we prompt an LLM-as-Judge to score model outputs. We validate this judge by reporting the inter-annotator agreement as Cohen's κ between the judge and human labels in the reference ground truth dataset. The LLM judge achieves at least fair agreement with human ground truth ($\kappa \geq 0.35$), with an average $\kappa = 0.46$, roughly comparable to the average human-human agreement in the ground truth datasets ($\kappa = 0.54$).

Domain	Feature	Classes	Ground Truth Dataset	Judge κ
User Simulators	Dialogue act	4	Daily Dialogue (Li et al., 2017)	0.555
	Emotion	7	Li et al. (2017)	0.371
	Emotional disclosure	2	Jaidka et al. (2020)	0.372
	Feedback type	6	Liu et al. (2025)	0.363
	Information disclosure	2	CL-Aff OffMyChest (Jaidka et al., 2020)	0.505
	Politeness strategy	4	Zhao & Hawkins (2025)	0.468
	Topic shift	2	TIAGE (Xie et al., 2021)	0.375
Patient Simulators	Approach–avoidance	4	Elliot (2006)	0.602
	Emotional arousal	3	Warwar & Greenberg (2000)	0.502
	Experiencing	3	Klein et al. (1969)	0.452
	Innovative moment	6	Gonçalves et al. (2011)	0.464
	Insight	4	Hill et al. (1992)	0.482
	Locus of control	4	Rotter (1966)	0.406
	Mentalising	4	Fonagy et al. (1998)	0.465
	Outcome expectancy	4	Devilly & Borkovec (2000)	0.415
	Problem presentation	4	Orlinsky (2009)	0.434
	Symptom expression	6	Clark & Watson (1991)	0.579

Table 1: **Feature basis for social simulation fidelity.** Compared to ground truth labels, our LLM-as-Judge achieves an average agreement of $\kappa = 0.46$, which is roughly comparable to the average human-human agreement in the ground truth datasets ($\kappa = 0.54$).

The user simulation features capture how real users respond to and steer interactions (Bernard & Balog, 2024). User simulators should establish realistic goals (Sekulić et al., 2024), ask reasonable questions (Yoon et al., 2024), maintain consistent preferences (Sekulić et al., 2024; Dou et al., 2025), and react with the appropriate emotions when their conversation partners behave inappropriately (Meshi et al., 2026). Moreover, simulated users should stress test agents by exhibiting impatience, topical digressions, and underspecified requests (Shim et al., 2026), requiring agents to elicit missing information and advance the conversation more proactively (Sekulić et al., 2022). Together, our seven user simulation features span the functional, discourse, interpersonal, and affective conversational states encountered by an assistant during training and evaluation.

The ten patient simulation features in the bottom half of Table 1 derive from the Generic Model of Psychotherapy (Orlinsky & Howard, 1986; Orlinsky, 2009; Norcross & Lambert, 2019), the Research Domain Criteria framework (Insel et al., 2010), and the Motivational Interviewing Treatment Integrity (MITI 4.2) (Moyers et al., 2016). Together, they measure whether the partner exhibits the relevant clinical stimuli by reconstructs the psychological, emotional, and relational signals to which the therapist responds.

2.3 Simulation Fidelity Metric

We define simulation fidelity distributionally as the divergence between the human and model feature distributions. These are conditional distributions, conditioned on both conversational history h and social context c defined in §2.1. Formally, let f denote a behavioral feature from §2.2, defined as a mapping from an utterance y to a one-hot feature vector. Summing over human utterances gives us

the human behavioral distribution $\pi_f^{\text{human}} = \Sigma f(y)$ and similarly, summing over LLM-generations $\hat{y} \sim P_m(y \mid h, c)$ gives us the model distribution $\pi_f^m = \Sigma f(\hat{y})$. The feature-level fidelity is defined as $D_f(m) = JSD(\pi_f^{\text{human}} \,\|\, \pi_f^m)$, the Jensen-Shannon divergence (Lin, 1991) between the human and the model distributions. Lower divergence indicates that the simulator reproduces the conditional frequency of the relevant human behavior more faithfully. Finally, for our downstream analyses, we report the average fidelity across features $\bar{D}(m) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}_s} D_f(m)$.

3 Current Gaps in Simulation Fidelity

Prior work suggests that frontier models can produce useful and stylistically plausible simulations (Zhu et al., 2026). However, because these models are trained to produce helpful and more cooperative language, they often fail to reproduce the same resistance that real human partners provide, both in therapy settings (Hoang et al., 2026) as well as in AI use (Naous et al., 2026). Therefore simulation fidelity is not always correlated with general task performance. In this section, we investigate the degree to which current models are faithful simulators of human partners.

Frontier models are unfaithful simulators. We evaluate three frontier models, *GPT 5.5* (OpenAI, 2026), *Claude Opus 5* (Anthropic, 2026), and *Gemini 3.1 Pro* (Google, 2026). As a point of comparison, we also evaluate the much smaller base model, *Delphi 2.5B* (Held, 2026) for which details of architecture and training data are available, and with which we can also study the effects of in-domain adaptation.[3] We define *Human-Level Fidelity* as the finite-sample divergence floor.

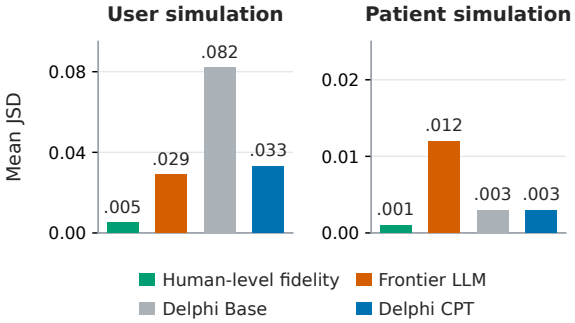


Figure 1: **Simulation fidelity results**[2] show a gap between frontier models and human partners.

The results in Figure 1 corroborate the prior literature, showing that frontier models are unfaithful simulators, and even under-perform much smaller base models. In *user simulation*, the best performing frontier model was Claude Opus 5. Although this model outperformed the smaller Delphi-2.5B Base model, there remains a large and significant gap separating its simulation from human levels of fidelity (0.029 vs 0.005). Similarly, in *patient simulation*, the best performing frontier model was Gemini, which achieved a simulation fidelity score $4\times$ worse than the much smaller Delphi Base model (JSD of 0.012 vs. 0.003).

Frontier models over-act and fail to provide resistance. We investigated the discrepancy across all three frontier models and found that, with respect to each our fidelity features, each model exhibits the same misrepresentation, and the same tendency to *over-act* and the same failure to provide *resistance*. Compared to both human patients and smaller base models, frontier model produce text with significantly higher levels of arousal and patient insight. Concretely, the *experiencing* scale of Klein et al. (1969) measures the degree to which the patient engages with their own inner experience. Compared to human levels (56%), models exhibit significantly higher levels of experiencing language (79% in GPT 5.5, 79% in Opus 5, 68% in Gemini 3.1 Pro). For example, in one transcript, the therapist asked a simple question about which bills the patient had, and the human patient simply listed them. The Delphi model also simply said *they're not too bad*, but all three frontier models

[2]Human-level fidelity is computed as the JSD noise floor. Frontier LLM is the best performing frontier model: Gemini 3.1 Pro for patient simulation and Claude Opus 5 for user simulation. Delphi CPT indicates the results of continued pre-training with Delphi 2.5B on 13.2M and 15M tokens of in-domain training data for patient and user simulation respectively.

[3]The Delphi suite contains models of the same size and architecture as today's behavioral foundation models like OdysSim (Zhou et al., 2026a) with the added benefit that Delphi is designed for scaling analysis, and Delphi's pre-training corpus, the DCLM, is fully known.

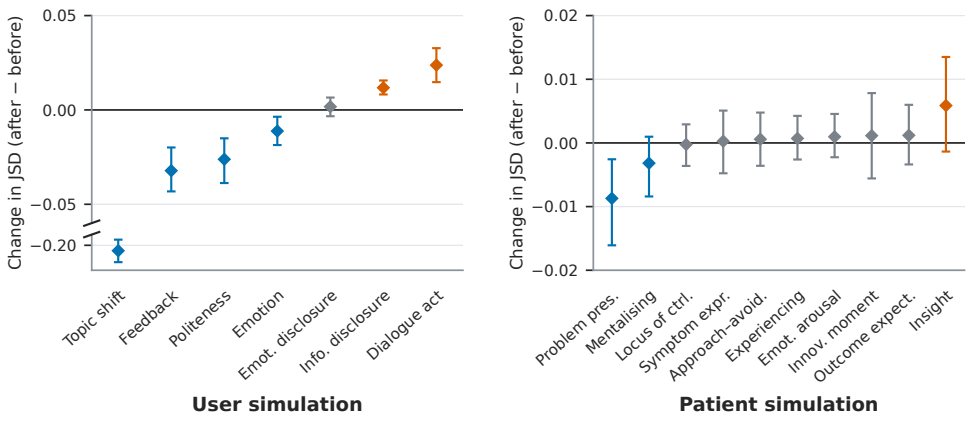


Figure 2: **Fine-tuning has heterogeneous effects on behavioral fidelity across simulation domains.** Negative values indicate movement toward the human distribution. Error bars show 95% bootstrap intervals.

embellished their reply by volunteering elaborate reflections like "*having a life–that's where I fall apart*" (Opus), or "*[I] feel embarrassed that this... counts as good*" (GPT 5.5).

Similarly, in user simulation, the frontier models provide low resistance by avoiding topic shifts (0.2% of turns in Claude vs. 9% in humans), and by asking more questions (74% in Claude vs. 45% in humans), which explicitly facilitate easier and more actionable responses from AI assistants. For example, one user asked an assistant where she should invest her money and the assistant recommended a financial advisor. The human user resisted the advice, asking "*how can I trust [him]?*" Claude Opus, on the other hand, accommodated and followed up with another easy question, "*Can you explain how index funds work?*"

Continued pre-training fails to close fidelity gaps. Fine-tuning can improve simulation fidelity (Zhu et al., 2026; Kolluri et al., 2025). We also investigate whether continued pre-training with a limited budget of available in-domain data improves simulation fidelity. To answer this, we perform continued pre-training (CPT) on Delphi-2.5B (Held, 2026) with the same causal language modeling objective, using 15M tokens of user interaction data from the OdysSim corpus (Zhou et al., 2026a), and 13M tokens of held-out psychotherapy transcripts from Alexander Street. Figure 1 shows that CPT improves user simulation fidelity by 2.5×, but the adapted model remains near the frontier model performance, still 7.2× above the human noise floor. For patient simulation, CPT produces no significant improvement, and the model still remains 3× above the human noise floor.

Figure 2 shows how the feature-specific changes introduced by continued pre-training are non-monotonic. In user simulation, four of the seven behavioral features move closer to the human distribution, while two features move away, and emotional disclosure shows no clear change. Patient simulation exhibits a similarly non-monotonic effect. Two features move closer to the human distribution, while the remaining eight show no clear improvement. We conclude that in-domain data can help improve model fidelity, but in our limited data regime (15M tokens), the benefits are mixed, and not great enough to recover human-level fidelity. This motivates our central question in §4, whether more general purpose and widely available pre-training can close the gap at scale.

4 Simulation Fidelity Scales with Pre-training Compute

After frontier model fidelity gaps and discovering that domain adaptation with limited data is not enough to close these gaps (§3), we now investigate web pre-training data as an alternative lever for improving performance. Webtext is a cheaper and more abundant alternative to in-domain data, so we test whether webtext might reduce the need for in-domain data at scale.

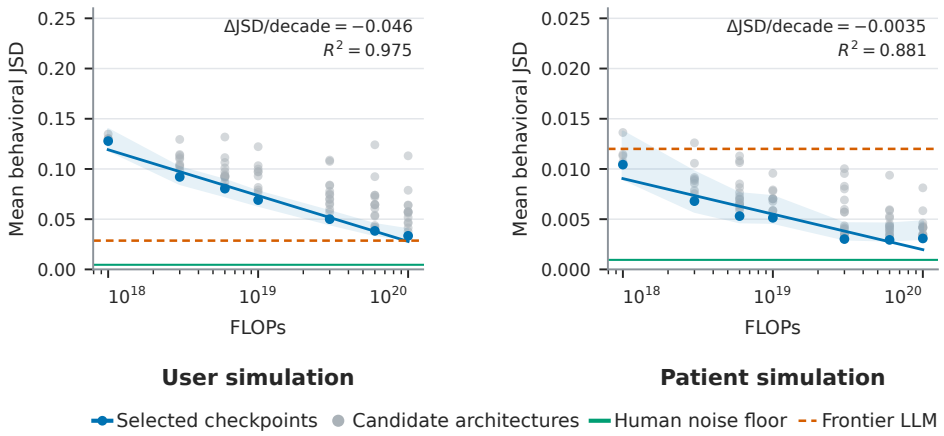


Figure 3: **Distributional fidelity scales log-linearly** for both **user simulation** and **patient simulation** with tight fits ($r^2 \geq 0.88$). Shading shows 90% bootstrap confidence intervals.

We formalize this investigation by fitting carefully-controlled compute scaling laws on the 85 parameter-controlled IsoFLOP checkpoints of Held et al. (2025). These models are decoder-only transformers pre-trainined on webtext from the DCLM corpus (Li et al., 2024) at seven compute budgets spanning 10^{18} to 10^{20} FLOPs. Following the methodology of Hoffmann et al. (2022), we fit our scaling laws to the performance of the best model at each compute budget. Here, best performance is defined by the lowest behavioral JSD score (§2.3).

Simulation fidelity is at least partly an emergent pre-training capability. Our results in Figure 3 show that fidelity improves in both simulation domains as a log-linear function of pre-training compute scale. These fits are tight, with $r^2 = 0.98$ for user simulation and $r^2 = 0.88$ for patient simulation respectively, indicating that, over two orders of magnitude, pre-training is a reliable lever for improving simulation fidelity. In user simulation, every $10\times$ increase in compute drops the JSD by 0.046, which closes as much as 37% of the initial gap to the human noise floor. Similarly, in patient simulation, the same increase in compute drops JSD by 0.0035, or up to 38% of the initial gap to the human noise floor. These results show that simulation fidelity appears increasingly available from webtext alone, in base models without any instruction tuning or explicit simulation objective.

At relatively small model scales, scaling complements domain adaptation with limited data. To achieve a performance benefit equivalent to domain adaptation with 15M tokens, one needs as little as $1.5\times$ the pre-training compute for user simulation. We conclude that, in smaller models, pre-training and in-domain adaptation are complementary levers, but especially where in-domain data is less abundant, pre-training can offer a scalable complement.

Post-training may have an insignificant or negative impact on simulation fidelity. One striking feature of Figure 3 is that small base models under 2.5B parameters rapidly approach frontier model performance in the case of user simulation, or even surpass it in the case of patient simulation. This echoes prior works and our own findings in §3 that large preference-tuned models exaggerate human behavior, introducing a negative impact on simulation fidelity. This result has significant implications for practitioners interested in running their own training. Instead of relying on frontier API-based models, our results suggest that practitioners should consider cheaper, smaller, and more local base models, which may be less prone to exaggeration, and which they can better inspect, intervene on, fine-tune, and reproduce.

The benefits of scale are unlikely to extrapolate. Although simulation fidelity improves predictably over the observed compute range (10^{18}-10^{20} FLOPs), these gains cannot extrapolate indefinitely. There is a nonzero fidelity floor, and already, we observe some evidence of diminishing returns at larger compute scales, despite the strong overall fits of the scaling laws. In patient simulation, we observe zero improvement in fidelity from the three largest model scales at 3×10^{19}, 6×10^{19}, 1×10^{20} FLOPs. This could suggest a performance asymptote due to residual errors that additional next-token prediction on webtext may not resolve. Similarly, in user simulation, the largest

model performs worse than the fitted scaling curve as it approaches the frontier performance of Opus 5. One possible implication is a regime change. Scale may be sufficient to reach the frontier, but it may not be sufficient to move beyond it. In this case, our results would further motivate research on simulation-specific methodologies. Simple interventions like persona conditioning are largely ineffective (Zhu et al., 2026), but other works have considered expert-derived cognitive constraints (Wang et al., 2024; Louie et al., 2024), new agent architectures (Park et al., 2024b), or reinforcement learning with simulation-specific rewards (Abdulhai et al., 2026; Wang et al., 2025).

5 Relative Scaling Laws for Simulation Fidelity

We have seen that the benefits of pre-training compute scale are significant, but these benefits are suggested mainly at smaller model scales (§4). This motivates us to further investigate whether the simulation domain benefits from scale as readily as other domains of text generation. If simulation fidelity does not keep pace with more general advancements in AI, then this could be a training bottleneck, and a new motivating reason to invest in alternative simulation-specific methodologies.

5.1 Language Modeling Loss as a Proxy for Simulation Fidelity

It is not possible to perform the cross-domain comparison of relative scaling laws with our domain-specific fidelity metrics. For this reason, we now consider whether language modeling loss can serve as a cheaper and more versatile proxy for simulation fidelity. Unlike our fidelity metric, language-modeling loss requires neither model sampling nor an expensive LLM judge for every output, making it better suited to the relative scaling analyses in §5. **Loss is a reasonable proxy for fidelity.** First we show that BPB and average behavioral JSD improve at similar rates with increasing compute. For each metric, we independently select the compute-optimal checkpoint at each of the seven compute budgets and order the resulting frontier values by compute, $C_1 < \cdots < C_7$. For each adjacent compute interval $i = 1, \ldots, 6$, we compute the improvement per log-unit of compute:

$$\text{BPB rate}_i = \frac{L_i - L_{i+1}}{\log_{10} C_{i+1} - \log_{10} C_i}, \qquad \text{JSD rate}_i = \frac{D_i - D_{i+1}}{\log_{10} C_{i+1} - \log_{10} C_i},$$

where L_i is the frontier validation BPB and D_i the average behavioral JSD at compute budget C_i. We quantify rate tracking as the Pearson correlation between six adjacent improvement rates:

$$r_{\text{rate}} = \text{Pearson}\big(\{\text{BPB rate}_i\}_{i=1}^{6}, \{\text{JSD rate}_i\}_{i=1}^{6}\big)\,.$$

We find rate-tracking correlations of $r_{\text{rate}} = 0.759$ for user simulation and $r_{\text{rate}} = 0.922$ for patient simulation, comparing the selected BPB and fidelity-optimal frontiers across the 85-model suite. These results suggest that loss tracks improvements in behavioral fidelity reasonably well.

Loss is an imperfect proxy for fidelity. Further investigation shows that language-modeling loss does not capture all aspects of behavioral fidelity, as subgroup-level analyses reveal discrepancies in the relative slopes which are not reflected by aggregate loss (Appendix A.1). Still, the correspondence in improvement rates supports conditional language-modeling loss as a practical proxy for overall simulation fidelity across compute. We use it in our relative-scaling analyses.

5.2 Relative Scaling Disparities

Methodology. We follow the *relative scaling laws* methodology of Held et al. (2025), which allows us to isolate the *rate* disparity from any initial performance disparity. Let $L_d^*(C)$ and $L_{\text{ref}}^*(C)$ denote the frontier losses for domain d and a reference domain at compute budget C. We define their relative loss as $G_d(C) = (L_d^*(C))/(L_{\text{ref}}^*(C))$ and fit its change with compute as

$$\log G_d(C) \approx \log \gamma_d + \Delta\beta_d \log(C/C_0).$$

Here, γ_d captures the relative loss at the initial compute budget C_0, whereas $\Delta\beta_d$ captures how that relative position changes as compute increases. For a neutral reference domain, we take the loss L_{code}^* computed on Python code. Again, we fit all scaling curves on the same 85 IsoFLOP checkpoints from §4 using five in-context examples.

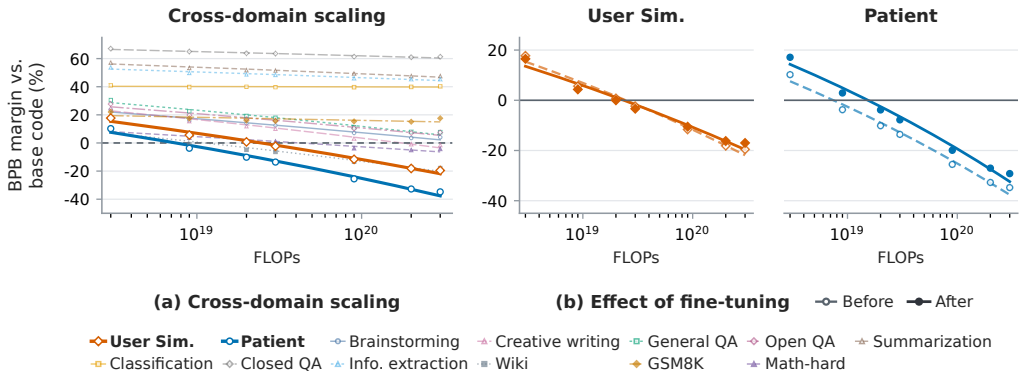


Figure 4: **Fine-tuning improves relative performance but does not eliminate social-simulation scaling disparities.** Both panels show BPB margins relative to code as a function of pre-training compute (FLOPs). **(a)** Before domain adaptation, social-simulation tasks are emphasized against a broad set of comparison tasks. **(b)** For user simulation (PRISM) and patient simulation, open markers and dashed lines denote performance before domain adaptation, while filled markers and solid lines denote performance after domain adaptation. Both stages use the unchanged base-model code BPB at each compute budget as their reference. Points denote evaluated checkpoints and lines denote fitted relative scaling laws. Declining trajectories indicate slower improvement than code as pre-training compute increases.

Simulation is the most relatively disadvantaged domain. In Figure 4 (a), we plot the relative loss margin $M_d(C) = 100\,(1 - L^*_d(C)/L^*_{\text{code}}(C))$ for our simulation tasks along with a wide range of tasks, including open-ended generation tasks like creative writing, brainstorming, and open QA. All tasks become easier to predict as compute increases, but their relative margins are markedly different. Unsurprisingly, more constrained and closed-form tasks like classification exhibit little marginal change relative to code. On the other hand, the relative margin for user simulation declines by 37.3 percentage points over the evaluated compute range. Patient simulation exhibits an even larger disparity, falling 45 points from 10.2% to -34.8%. These disparities are significantly more severe than what we observe from any other open-ended generation domain. The results suggest that simulation will become a relative bottleneck as general performance outpaces simulation fidelity, providing another motivation for AI research to invest in new simulation methodologies.

5.3 In Domain Data Does Not Mitigate Relative Disparities

One explanation for slower relative is *domain underexposure*. Realistic user interaction and psychotherapy transcripts may be comparatively scarce in web-scale pre-training corpora, such that additional compute mostly exposes models to other forms of text. If this is the case, we might expect that domain adaptation could amplify the return obtained from additional pre-training compute (Magnusson et al., 2024). The alternative is that exposure under our limited data regime is insufficient. In this case, models pre-trained with more compute may not derive greater improvements in behavioral fidelity, even when limited additional in-domain text is available.

Domain-adaptive pre-training provides a way to distinguish these possibilities while holding the training objective fixed (Gururangan et al., 2020). We distinguish between two effects. Adaptation may change the *level* of the scaling curve, improving performance at each pre-training scale, or it may change the *rate* of scaling, altering how much additional performance is obtained from increasing pre-training compute. If in-domain scarcity is responsible for the relative scaling disparity, additional domain exposure might steepen the scaling trajectory and narrow that disparity. For both domains, we continue pre-training seven matched Delphi checkpoints spanning 3×10^{18} to 3×10^{20} pre-training FLOPs using the causal language-modeling objective, following the methodology in §3. Full training details are provided in Appendix D.

Domain adaptation does not consistently improve the scaling rate. Figure 4 (b) shows the same pattern across simulation domains: domain adaptation improves performance at a given compute budget, but the relative marginal slopes follow the same negative trends. We find no evidence that in-domain adaptation improves the relative rate. In patient simulation, the estimated change in the

scaling coefficient is only $\Delta\beta = 0.020$, with a 95% confidence interval that includes zero. We conclude that, although continued pre-training on can improve simulation performance, in such a limited data regime it does not fix relative scaling disparities. For this reason, we recommend greater investment in larger-scale in-domain datasets than what are currently available.

6 Related Work

Evaluating simulation fidelity. Evaluations measure distinct properties. Human raters and LLM judges have assessed the naturalness, contextual coherence, and perceived realism of role-play dialogue (Lu et al., 2025). Downstream evaluations measure whether simulated users preserve assistant rankings (Dou et al., 2025), improve agent's performance in subsequent interactions with real users (Suh et al., 2026), or support practitioner skill gains when paired with feedback (Louie et al., 2026). Behavioral fidelity asks how closely simulated responses reproduce human responses under comparable conditions. Simulated and human patients, for example, can discuss similar topics while differing in their expression patterns (Hoang et al., 2026); evaluations of simulated students likewise distinguish linguistic and cognitive dimensions (Scarlatos et al., 2026). We therefore validate LLM judge annotations of domain-relevant behaviors against human annotations, then compare simulated and human feature distributions.

Building more faithful LLM simulations of training partners. Simulation fidelity is not correlated with the model's general task performance (Naous et al., 2026; Tjuatja et al., 2024), and persona prompting alone is not a straightforward mitigation (Zhu et al., 2026). Policies trained with such unfaithful simulators exhibit poor generalization in real-world deployments (Yu et al., 2026). A more effective method is to ground generation in rich behavioral profiles derived from in-depth human interviews (Park et al., 2024a) or expert-derived cognitive constraints (Wang et al., 2024; Louie et al., 2024), but this is time consuming. Additionally, with enough in-domain interaction trajectories, it is possible to fine-tune a faithful simulator (Zhu et al., 2026; Kolluri et al., 2025). However, such domain-specific adaptation is often slow, expensive, and difficult to scale: the same barriers that restrict training with human partners.

Scaling laws. The closest scaling analysis of simulation performance evaluates MCQ tasks like opinion modeling, behavioral prediction, and longitudinal forecasting (Ziems et al., 2026). We extend this question to **open-ended, next-turn responses** from simulated users and therapy patients, measuring fidelity against human behavioral distributions.

7 Conclusion

In this work, we proposed new metrics to measure the fidelity of LLM simulations for partner training and found evidence of persistent performance gaps in frontier models, which could not be fully mitigated by in-domain adaptation. This motivated our investigation of compute scaling laws, where we demonstrated the predictable benefit of pre-training, as well as its limitations. Pre-training alone appears insufficient to close the performance gap. To conclude, we recommend that AI researchers invest more research attention towards the problem of simulation fidelity.

Limitations and future work. Behavioral fidelity estimates rely on LLM judges whose agreement with human annotations varies across features and is lower on text from the smallest checkpoints. Feature-level scaling estimates are therefore less reliable at the low-compute end of the range. More broadly, if additional in-domain data does not consistently change the relative scaling rate, the question is what would. One possibility is the adaptation budget: every base model receives the same absolute number of adaptation tokens, so larger models receive progressively less adaptation relative to pre-training. Another is data coverage: behaviors left unchanged by adaptation may be underrepresented in the training corpus. A third is the training objective, since next-token prediction optimizes likelihood rather than agreement with human behavioral distributions. The combination of a large reduction in language-modeling loss and selective behavioral gains is consistent with this possibility. Future work can distinguish these explanations by scaling the adaptation budget with model size, using broader in-domain corpora with better coverage of underrepresented behaviors, and directly optimizing distributional fidelity under matched compute.

AI USE STATEMENT

During manuscript preparation, AI tools were used to identify relevant literature, search for information, and prepare tables and figures. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies the scaling behavior of LLM-based human simulators. More faithful simulators may benefit education, clinical training, and human–AI collaboration, but they may also enable misuse through impersonation, manipulation, or deceptive interactions. Our fidelity metrics capture selected behavioral distributions and should not be interpreted as establishing comprehensive fidelity to human behavior or as a substitute for evaluation with real people in high-stakes settings.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the simulation tasks, datasets, prompts, model configurations, and fidelity metrics used in our experiments. All datasets used in this work are accessible for research use, and we report the sampling and evaluation procedures needed to reconstruct our experimental sets. We will release our evaluation code, prompts, and experiment configurations to facilitate reproduction of the reported results.

REFERENCES

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *Advances in Neural Information Processing Systems*, 38:52920–52957, 2026.

Alexander Street. Counseling and psychotherapy transcripts: Volume I. Alexander Street Press, Alexandria, VA (now part of Clarivate). Licensed database, accessed 2026, 2008. URL https://search.alexanderstreet.com/psyc.

Alexander Street Press. Counseling and psychotherapy transcripts: Volume i. Database and streaming collection, 2013. URL https://search.alexanderstreet.com/psyc.

Anthropic. Claude opus 5 api. Anthropic API, 2026. URL https://www.anthropic.com.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Nolwenn Bernard and Krisztian Balog. Towards a formal characterization of user simulation objectives in conversational information access. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 185–193, 2024.

Harshita Chopra, Kshitish Ghate, Aylin Caliskan, Tadayoshi Kohno, Chirag Shah, and Natasha Jaques. Beyond cooperative simulators: Generating realistic user personas for robust evaluation of llm agents. *arXiv preprint arXiv:2605.12894*, 2026.

Lee Anna Clark and David Watson. Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology*, 100(3):316–336, 1991. doi: 10.1037/0021-843X.100.3.316.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Grant J. Devilly and Thomas D. Borkovec. Psychometric properties of the credibility/expectancy questionnaire. *Journal of Behavior Therapy and Experimental Psychiatry*, 31(2):73–86, 2000. doi: 10.1016/S0005-7916(00)00012-4.

Kate O Donovan, Michael Connolly, and Freda Browne. Within healthcare education how do simulated patients enable (or not) the transfer of learning into clinical practice: A realist review. *Nurse Education Today*, pp. 106927, 2025.

Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of ai assistants? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35200–35278, 2025.

Andrew J. Elliot. The hierarchical model of approach-avoidance motivation. *Motivation and Emotion*, 30(2):111–116, 2006. doi: 10.1007/s11031-006-9028-7.

Peter Fonagy, Mary Target, Howard Steele, and Miriam Steele. Reflective-functioning manual version 5 for application to adult attachment interviews. Unpublished manuscript, University College London, 1998.

Miguel M. Gonçalves, António P. Ribeiro, Inês Mendes, Marlene Matos, and Anita Santos. Tracking novelties in psychotherapy process research: The innovative moments coding system. *Psychotherapy Research*, 21(5):497–509, 2011. doi: 10.1080/10503307.2011.560207.

Google. Gemini 3.1. Large language model, 2026. URL https://gemini.google.com. Accessed: 2026-09-25.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Will Held. Scaling laws that extrapolate 300× past the fit. https://www.openathena.ai/blog/delphi/, may 2026. Open Athena Blog.

William Held, David Hall, Percy Liang, and Diyi Yang. Relative scaling laws for llms. *arXiv preprint arXiv:2510.24626*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Clara E. Hill, Maureen M. Corbett, Bettina Kanitz, Peggy Rios, Richard Lightsey, and Maria Gomez. Client behavior in counseling and therapy sessions: Development of a pantheoretical measure. *Journal of Counseling Psychology*, 39(4):539–549, 1992. doi: 10.1037/0022-0167.39.4.539.

Van Hoang, Eoin Rogers, and Robert Ross. LLMs as standardised patients for motivational interviewing: How faithful are they? In Aya Zirikly, Kfir Bar, Sean MacAvaney, Molly Ireland, Yaakov Ophir, Dana Atzil-Slonim, Vasudha Varadarajan, Steven Bedrick, and Bart Desmet (eds.), *Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026)*, pp. 258–270, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-421-7. doi: 10.18653/v1/2026.clpsych-1.21. URL https://aclanthology.org/2026.clpsych-1.21/.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S. Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7): 748–751, 2010.

Kokil Jaidka, Iknoor Singh, Jiahui Lu, Niyati Chhaya, and Lyle Ungar. A report of the CL-Aff OffMyChest Shared Task: Modeling Supportiveness and Disclosure. In *Proceedings of the AAAI-20 Workshop on Affective Content Analysis*, New York, USA, 2020. AAAI.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024. URL https://arxiv.org/abs/2404.16019.

Marjorie H. Klein, Philippa L. Mathieu, Eugene T. Gendlin, and Donald J. Kiesler. *The Experiencing Scale: A Research and Training Manual*, volume 1. Wisconsin Psychiatric Institute, Madison, WI, 1969.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. Finetuning llms for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025.

Deuksin Kwon, Kaleen Shrestha, Bin Han, Spencer Lin, James Hale, Jonathan Gratch, Maja Mataric, and Gale M Lucas. Can llms truly embody human personality? analyzing ai and human behavior alignment in dispute resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38790–38798, 2026.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Kamal Mohamed Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Joshua P Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander T Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=CNWdWn47IE.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset, 2017. URL https://arxiv.org/abs/1710.03957.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Yi Liu, TingFeng Hui, Wei Zhang, Li Sun, Ningxin Su, Jian Wang, and Sen Su. Envsimbench: A benchmark for evaluating and improving llm-based environment simulation. *arXiv preprint arXiv:2605.07247*, 2026.

Yuhan Liu, Michael JQ Zhang, and Eunsol Choi. User feedback in human-LLM dialogues: A lens to understand users but noisy as a learning signal. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2666–2681, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.133. URL https://aclanthology.org/2025.emnlp-main.133/.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10570–10603, 2024.

Ryan Louie, Raj Sanjay Shah, Ifdita Hasan Orney, Juan Pablo Pacheco, Emma Brunskill, and Diyi Yang. Can llm-simulated practice and feedback upskill human counselors? a randomized study with 90+ novice counselors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–31, 2026.

Dongxu Lu, Johan Jeuring, and Albert Gatt. Evaluating LLM-generated versus human-authored responses in role-play dialogues. In Lucie Flek, Shashi Narayan, Lê H`ông Phuong, and Jiahuan Pei (eds.), *Proceedings of the 18th International Natural Language Generation Conference*, pp. 20–40, Hanoi, Vietnam, October 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.inlg-main.2/.

Rebecca Luebbert, Amelia Perez, Angela Andrews, and Tracy Webster-Cooley. Standardized patients versus mannequins in mental health simulation. *Journal of the American Psychiatric Nurses Association*, 29(4):283–289, 2023.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. Can llms effectively simulate human learners? teachers' insights from tutoring llm students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 100–117, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

Ofer Meshi, Krisztian Balog, Sally Goldman, Avi Caciularu, Guy Tennenholtz, Jihwan Jeong, Amir Globerson, and Craig Boutilier. Convapparel: A benchmark dataset and validation framework for user simulators in conversational recommenders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5270–5304, 2026.

Theresa B. Moyers, Lauren N. Rowell, Jennifer K. Manuel, Denise Ernst, and Jon M. Houck. The motivational interviewing treatment integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42, 2016. doi: 10.1016/j.jsat.2016.01.001.

Yannik Müller, Hans Lemberg, Paul Gellert, and Jan C Zoellick. Training empathetic communication skills in medical students with a role-prompted gpt-4o chatbot: quasi-experimental intervention study. *JMIR Medical Education*, 12:e88092, 2026.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. Flipping the dialogue: Training and evaluating user language models. In *International Conference on Learning Representations*, volume 2026, pp. 82079–82103, 2026.

Ha Nguyen and Jie Cao. Exploring the capacity of large language models to simulate students' scientific thinking: Insights for responsive teaching. *Journal of Science Education and Technology*, pp. 1–18, 2026.

John C. Norcross and Michael J. Lambert (eds.). *Psychotherapy Relationships That Work, Volume 1: Evidence-Based Therapist Contributions*. Oxford University Press, New York, 3 edition, 2019.

OpenAI. GPT-5.5 System Card. Technical report, OpenAI, 4 2026. URL https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf.

David E. Orlinsky. The "generic model of psychotherapy" after 25 years: Evolution of a research-based metatheory. *Journal of Psychotherapy Integration*, 19(4):319–339, 2009. doi: 10.1037/a0017973.

David E. Orlinsky and Kenneth I. Howard. Process and outcome in psychotherapy. In Sol L. Garfield and Allen E. Bergin (eds.), *Handbook of Psychotherapy and Behavior Change*, pp. 311–381. Wiley, New York, 3 edition, 1986.

Sitong Pan, Robin Schmucker, Bernardo Garcia Bulle Bueno, Salome Aguilar Llanes, Fernanda Albo Alarcón, Hangxiao Zhu, Adam Teo, and Meng Xia. Tutorup: What if your students were simulated? training tutors to address engagement challenges in online learning. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–18, 2025.

Joon Sung Park, Carolyn Q Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, et al. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024a.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024b.

Akshara Prabhakar, Zuxin Liu, Ming Zhu, Jianguo Zhang, Tulika Manoj Awalgaonkar, Shiyu Wang, Zhiwei Liu, Haolin Chen, Thai Hoang, Juan Carlos Niebles, et al. Apigen-mt: Agentic pipeline for multi-turn data generation via simulated agent-human interplay. *Advances in Neural Information Processing Systems*, 38, 2026.

Julian B. Rotter. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1):1–28, 1966. doi: 10.1037/h0092976.

Alexander Scarlatos, Ryan S Baker, and Andrew Lan. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th international learning analytics and knowledge conference*, pp. 249–259, 2025.

Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. Simulated students in tutoring dialogues: Substance or illusion? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42349–42385, 2026.

Ivan Sekulić, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. Reliable llm-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pp. 19–35, 2024.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pp. 888–896. ACM, February 2022. doi: 10.1145/3488560.3498440. URL http://dx.doi.org/10.1145/3488560.3498440.

Jeonghoon Shim, Woojung Song, Cheyon Jin, Seungwon KooK, and Yohan Jo. Non-collaborative user simulators for tool agents, 2026. URL https://arxiv.org/abs/2509.23124.

Joseph Suh, Ayush Raj, Minwoo Kang, and Serina Chang. Quantifying the utility of user simulators for building collaborative llm assistants, 2026. URL https://arxiv.org/abs/2605.09808.

Weiwei Sun, Xuhui Zhou, Weihua Du, Xingyao Wang, Sean Welleck, Graham Neubig, Maarten Sap, and Yiming Yang. Training proactive and personalized llm agents. *arXiv preprint arXiv:2511.02208*, 2025.

Xihe Tian, Yanhong Liu, Keni Li Zhang, Sarah Litt, Kahyen Shin, and Tanqiu Jiang. Do artificial intelligence clients speak like human clients? exploring gpt-4's content-level performance in counseling role-play. *Journal of Counseling & Development*, 104(1):14–26, 2026.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Jimmy Kyaw Tun, Guillaume Alinier, Jessica Tang, and Roger L Kneebone. Redefining simulation fidelity for healthcare education. *Simulation & Gaming*, 46(2):159–174, 2015.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21082–21107, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1025. URL https://aclanthology.org/2025.acl-long.1025/.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate Hardy, Hong Shen, Fei Fang, and Zhiyu Zoey Chen. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2405.19660.

Serine Warwar and Leslie S. Greenberg. Client emotional arousal scale-iii. Unpublished manuscript, York University, Toronto, 2000.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-MI: A dataset of expert-annotated counselling dialogues. In *ICASSP*, 2022.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1684–1690, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.145. URL https://aclanthology.org/2021.findings-emnlp.145/.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*, 2024.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1490–1504, 2024.

Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Jie Sun, Xiang Li, Jingxian He, Wenyue Hua, et al. Simulated patient systems powered by large language model-based ai agents offer potential for transforming medical education. *Communications Medicine*, 6(1):27, 2025.

Simon Yu, Nicholas Tomlin, Marwa Abdulhai, Ximing Lu, Derek Chong, Abe Hou, Dilara Soylu, Sergey Levine, Christopher D Manning, and Weiyan Shi. One frozen simulator is not enough: Simulator collapse in multi-agent rl. *arXiv preprint arXiv:2608.12253*, 2026.

Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4NtoAVqfhA.

Haoran Zhao and Robert D. Hawkins. Comparing human and LLM politeness strategies in free production. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16188–16216, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.820. URL https://aclanthology.org/2025.emnlp-main.820/.

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

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, et al. Mind the sim2real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026b.

Ming Zhu, Juntao Tan, Rithesh Murthy, Jielin Qiu, Liangwei Yang, Wenting Zhao, Silvio Savarese, Shelby Heinecke, and Huan Wang. Realusersim: Bridging the reality gap in agent benchmarking via grounded user simulation. *arXiv preprint arXiv:2605.20204*, 2026.

Caleb Ziems, William Held, Karaca, Max Grusky, Tatsunori Hashimoto, and Diyi Yang. Will scaling improve social simulation with LLMs? *arXiv preprint arXiv:2607.02464*, 2026. URL https://arxiv.org/abs/2607.02464.

A Appendix

A.1 Scaling across respondent subgroups

Figures 5–8 report scaling trajectories across all retained PRISM respondent and conversational characteristics. Within each characteristic, categories are matched to the size of the smallest retained category, producing matched cohorts ranging from 30 to 196 respondents per category. On average, smaller matched cohorts have wider bootstrap uncertainty intervals and may therefore yield less precisely estimated behavioral-JSD trajectories. However, cohort size is confounded with subgroup definition and category composition, so differences in fit tightness cannot be attributed to sample size alone.

Conditional BPB is shown on the left and average behavioral JSD on the right. Each metric independently selects its best-performing full-sample checkpoint at each of the seven compute budgets from the complete 85-checkpoint IsoFLOP pool, and all subgroup categories reuse the corresponding metric-specific checkpoints. We fit BPB as $\log_{10} L = \alpha + s_{\mathrm{BPB}} \log_{10} C$ and mean behavioral JSD linearly in log compute, $\bar{D} = a + b \log_{10} C$. Thus, s_{BPB} denotes the BPB log–log slope, while b denotes the change in mean JSD per $10\times$ increase in pre-training compute. These panels are descriptive and do not, by themselves, test differences between categories.

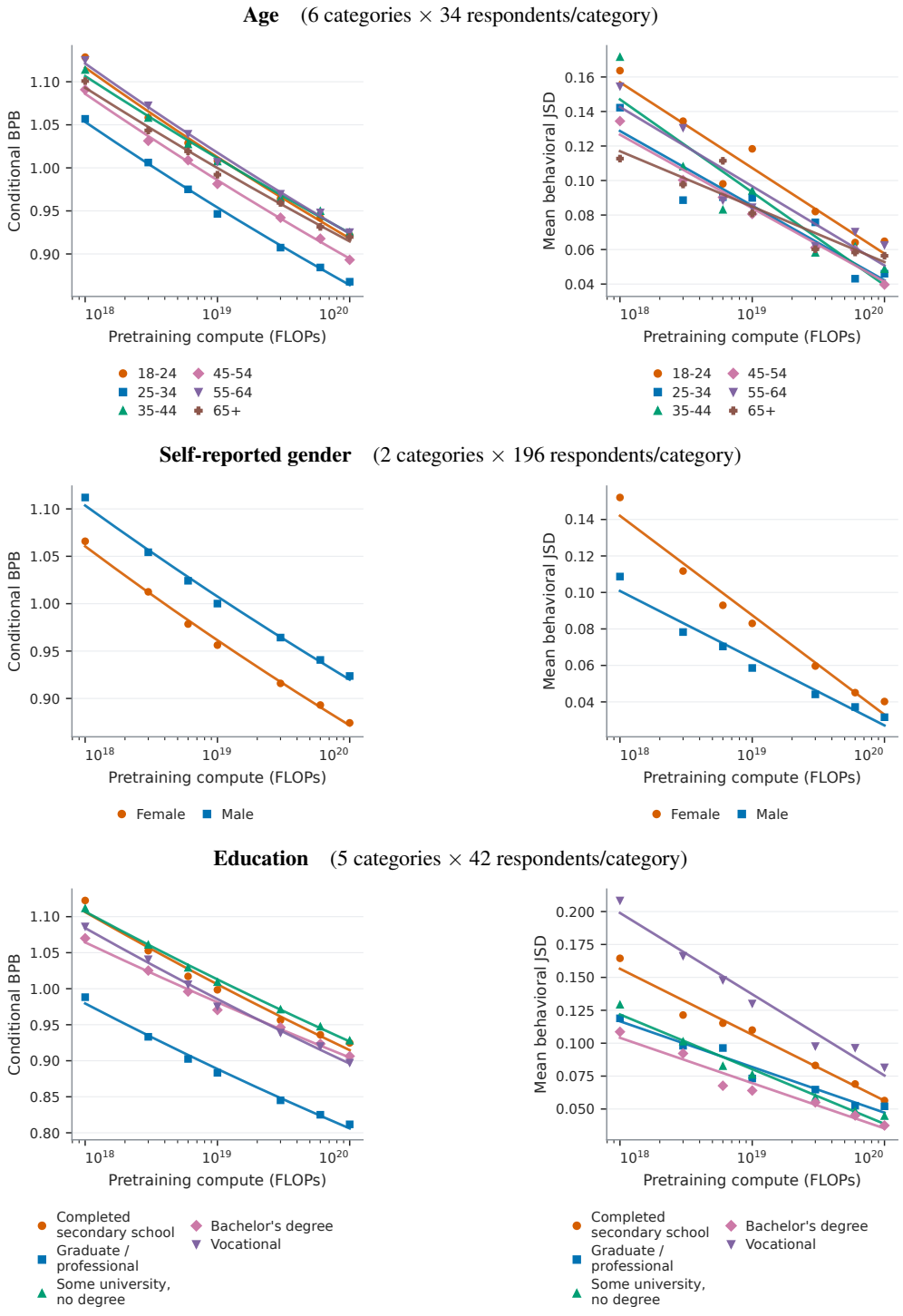


Figure 5: **Subgroup scaling results, Part I: age, gender, and education.** Conditional BPB is shown on the left and mean behavioral JSD on the right. Parenthetical counts give the number of retained categories and matched respondents per category. Across the 13 category-level trajectories, the median BPB log–log slope is $s_{\mathrm{BPB}} = -0.0414$ (range: -0.0431 to -0.0353), with median $R^2 = 0.994$ (range: 0.981–0.998). The median mean-JSD slope is -0.0433 per $10\times$ increase in compute (range: -0.0618 to -0.0321), with median $R^2 = 0.959$ (range: 0.854–0.975). Lower BPB and JSD indicate better predictability and behavioral fidelity, respectively.

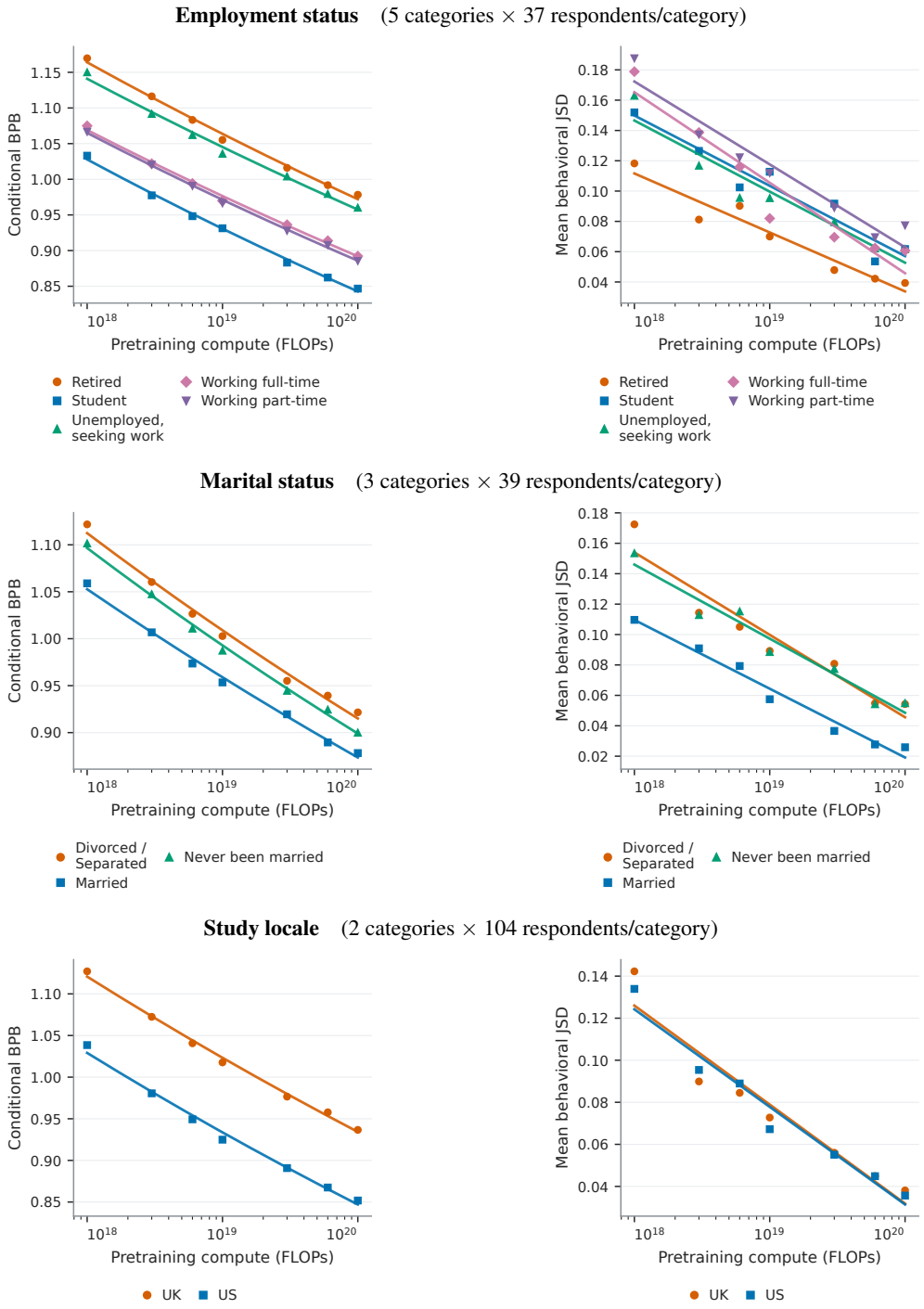


Figure 6: **Subgroup scaling results, Part II: employment, marital status, and study locale.** Conditional BPB is shown on the left and mean behavioral JSD on the right. Parenthetical counts give the number of retained categories and matched respondents per category. Across the 10 category-level trajectories, the median BPB log–log slope is $s_{\mathrm{BPB}} = -0.0403$ (range: -0.0431 to -0.0381), with median $R^2 = 0.995$ (range: 0.991–0.998). The median mean-JSD slope is -0.0470 per $10\times$ increase in compute (range: -0.0597 to -0.0389), with median $R^2 = 0.929$ (range: 0.915–0.975).

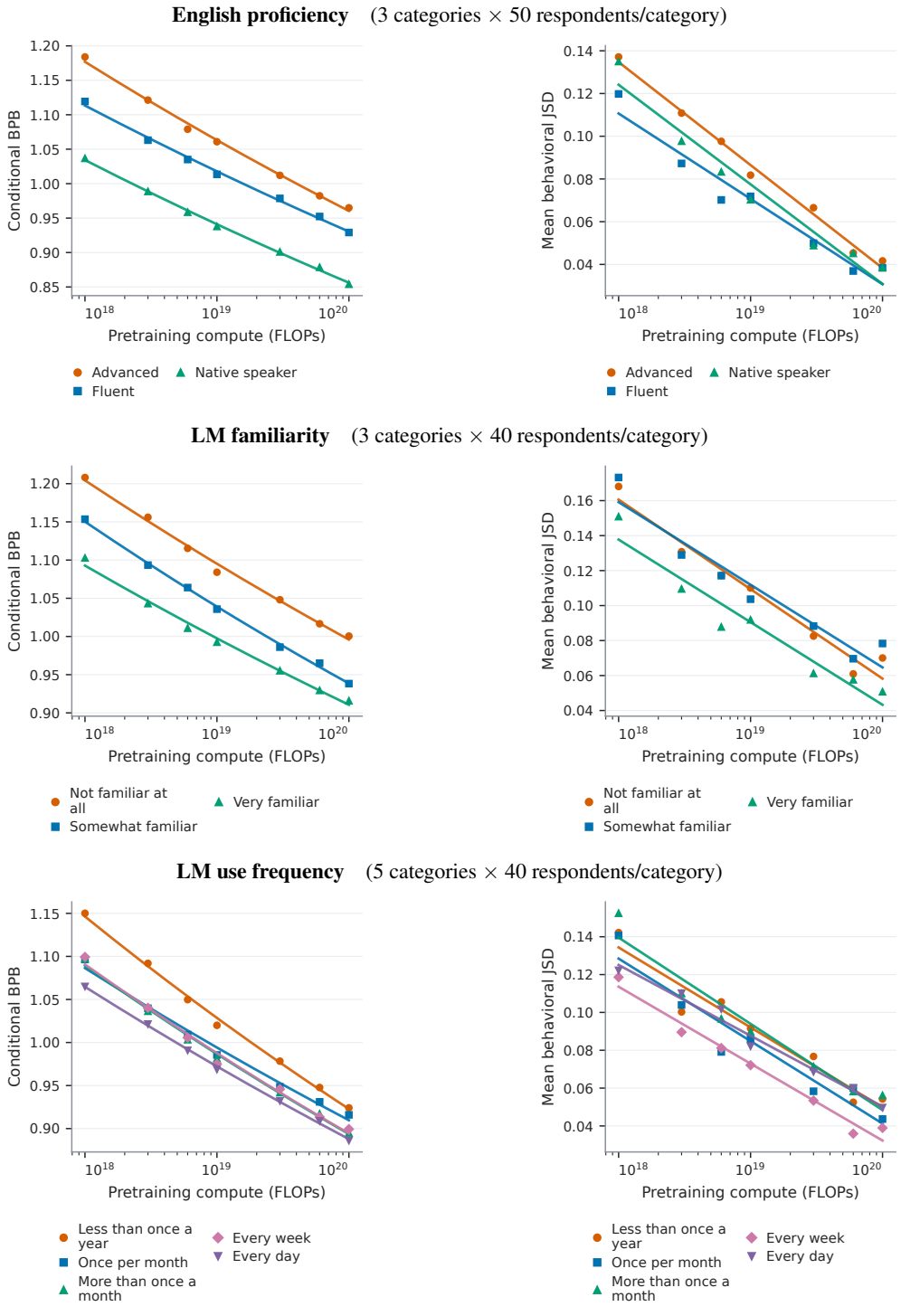


Figure 7: **Subgroup scaling results, Part III: language and language-model experience.** Conditional BPB is shown on the left and mean behavioral JSD on the right. Parenthetical counts give the number of retained categories and matched respondents per category. Across the 11 category-level trajectories, the median BPB log–log slope is $s_{\mathrm{BPB}} = -0.0413$ (range: -0.0471 to -0.0385), with median $R^2 = 0.996$ (range: 0.989–0.999). The median mean-JSD slope is -0.0456 per $10\times$ increase in compute (range: -0.0512 to -0.0374), with median $R^2 = 0.950$ (range: 0.919–0.991).

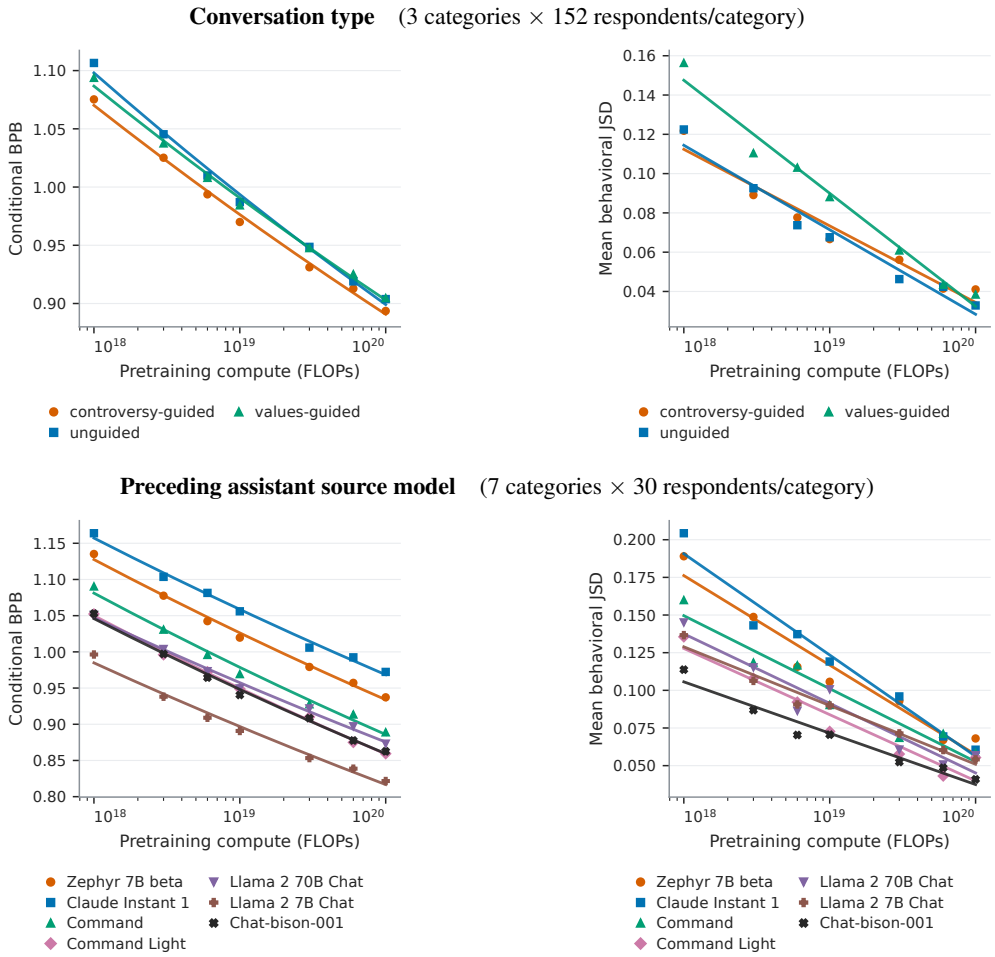


Figure 8: **Subgroup scaling results, Part IV: conversational characteristics.** Conditional BPB is shown on the left and mean behavioral JSD on the right. Parenthetical counts give the number of retained categories and matched respondents per category. Across the 10 category-level trajectories, the median BPB log–log slope is $s_{\mathrm{BPB}} = -0.0408$ (range: -0.0436 to -0.0387), with median $R^2 = 0.995$ (range: 0.987–0.998). The median mean-JSD slope is -0.0451 per $10\times$ increase in compute (range: -0.0673 to -0.0341), with median $R^2 = 0.954$ (range: 0.916–0.979).

B Cross-domain scaling

Databricks Dolly. We construct eight instruction-following control tasks from the Databricks Dolly corpus (Conover et al., 2023): general question answering, open question answering, closed question answering, brainstorming, classification, information extraction, summarization, and creative writing. An example is eligible when its instruction and response are non-empty. The prompt preserves the optional reference context when one is provided, and the target is the original human-written response. We do not apply additional date, demographic, or response-length filters.

Each control task is sampled independently and contains five fixed in-context demonstrations and 95 disjoint test examples. Categories are evaluated separately rather than pooled, and no validation split is used. As with PRISM, the split named `train` contains only the demonstrations and is not a parameter-training set.

Alexander Street psychotherapy transcripts We construct two next-turn prediction tasks from the Counseling and Psychotherapy Transcripts collection (Alexander Street, 2008): continuing as the patient and continuing as the therapist. A transcript is eligible when both speakers are present and it has at least 20 turns. A target turn is eligible when it is spoken by the target role, has at least 120

characters, and is preceded by six turns that include the other speaker. The prompt is a one-sentence persona built from the session metadata (age range, gender, marital status, presenting symptom, school of therapy), the six preceding turns, and the role tag. Each task is sampled independently and contains 500 examples, drawn round-robin over the 65 presenting-symptom groups with a fixed seed. Each example is scored with five demonstrations drawn with a fixed seed from the other examples in the same task.

Reference corpora Relative scaling requires reference domains evaluated under the same protocol as the psychotherapy tasks. We therefore score four public corpora with the identical five-shot next-continuation setup, 500 documents each. The reference is code: MBPP (Austin et al., 2021), where the model continues a task description with the reference Python program.[4] Mathematics enters at two difficulty levels, grade-school word problems from GSM8K (Cobbe et al., 2021) and competition problems from MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024), in both cases continuing the problem with its worked solution.[5] Finally, WikiText-103 (Merity et al., 2017) provides general expository prose: given one paragraph of an encyclopedia article, the model predicts the following paragraph.[6]

C Prompts and Generation Settings

C.1 User Simulation

PRISM. We construct the user-simulation evaluation set from the PRISM Alignment Dataset (Kirk et al., 2024), which contains conversations in which human participants interacted with responses from multiple language models and selected a preferred response. Each example is formulated as a next-user-turn prediction problem: given a participant message u_t and the assistant response a_t selected by that participant, the simulator generates the participant's subsequent message u_{t+1}. Rejected assistant responses are not included. Self-reported participant characteristics are retained only as analysis metadata and are never shown to the simulator.

We retain transitions for which the current and subsequent interaction turns each contain exactly one human message and the current turn contains exactly one selected assistant response. The participant message, selected response, and target message must be non-empty, written in English, and unflagged for personally identifiable information. We additionally require the target to contain 30–200 words and to fit within the generation budget, while the full few-shot prompt must fit within the models' context window. We remove examples with target leakage or inconsistencies between the reconstructed conversation and the source records.

Sampling is performed over unique participants so that participants who contributed more turns are not more likely to be selected. The resulting test set contains 456 examples from 456 participants, balanced across the three PRISM conversation types (152 examples per type). Five fixed in-context demonstrations are drawn from additional participants, with no participant or conversation overlap between the demonstration and test sets. For subgroup analyses, we construct equal-size cohorts separately for each retained level of age, gender, study locale, language-model familiarity and use frequency, education, employment status, marital status, English proficiency, conversation type, and preceding assistant model. These cohorts control the number of participants compared within each attribute, but do not impose joint balance across intersections of attributes. The split named `train` contains only the in-context demonstrations and is not used to update model parameters. Each demonstration is formatted in the same conversational template used at evaluation time, ending with the participant's actual subsequent response, e.g.,

```
Conversation:
Human: [participant message]

Assistant: [selected assistant response]
```

[4] https://huggingface.co/datasets/google-research-datasets/mbpp

[5] https://huggingface.co/datasets/openai/gsm8k, https://huggingface.co/datasets/HuggingFaceH4/MATH-500

[6] https://huggingface.co/datasets/Salesforce/wikitext

```
Human: [actual next participant response]
```

C.1.1 GENERATION

For each PRISM test example, we prompt the simulator with the five fixed in-context demonstrations described above, followed by the held-out conversation with the final `Human:` turn left for completion. In the demonstrations, this final turn contains the participant's actual subsequent response; in the test example, the same position is left empty for the model to generate. The generation instruction is:

```
Continue the conversation by writing exactly one next
message from the human participant. Return only that
message; do not include a speaker label, explanation, or
additional turn.
```

Here, "human participant" refers to the conversational role demonstrated by the five in-context examples. Each example consists of a complete multi-turn conversation between a human participant and an assistant, ending with the participant's actual subsequent response. These five complete conversations are provided as in-context demonstrations before target conversation. The "human participant" instruction reinforces the role illustrated by these examples, encouraging the model to generate a continuation that resembles how a human participant speaks and interacts rather than defaulting to an assistant-style response. Finally, for each test context, we sample five unique responses to account for stochastic variation under temperature 1.0; each response is generated with a maximum length of 256 tokens.

C.1.2 PRISM BEHAVIORAL ANNOTATION PROMPTS

Each of the seven behavioral features is coded independently using a separate feature-specific prompt and categorical label space. The same LLM-based judge, utilizing GPT 4.1-mini, is applied to both human-written and model-generated responses. Here, "human" denotes the user-side conversational role rather than the source of the response; model-generated continuations occupy this same role during evaluation. We use the same role framing for both response types so that the annotation prompt does not explicitly reveal whether a response is human-written or model-generated.

Each prompt produces only the label for its designated feature; labels from one feature are not provided as inputs to any other feature. All seven prompts receive the same two inputs:

- `context`: "Conversation containing the human question and assistant response."
- `candidate_response`: "Candidate genuine next message from the human."

C.1.3 DIALOGUE ACT

Prompt. Classify the candidate's primary communication intention in context. Identify the social action performed by the utterance, not its grammatical form, topic, correctness, or quality. Judge only this construct; do not infer any other behavioral dimension. Choose exactly one dialogue-act label.

Output field: `dialogue_act` (integer). The dialogue act is the candidate's main communicative function.

`0 = inform` The speaker contributes information they present as part of the common ground, including an answer, statement, explanation, opinion, confirmation, or denial.

`1 = question` The speaker seeks information from the addressee.

`2 = directive` The speaker tries to shape what the addressee does, as in a request, instruction, or suggestion.

`3 = commissive` The speaker undertakes, offers, accepts, or refuses a course of action for themself.

Classify the utterance's intended conversational function even when its sentence form is indirect or serves more than one minor purpose.

C.1.4 FEEDBACK TYPE

Prompt. Classify the candidate's relationship to the immediately preceding assistant reply. Feedback is backward-looking information about whether that reply met the user's intent or how it should be repaired; a turn that only advances the conversation is not feedback. Judge this relationship in context, not from keywords or sentence form, and choose one label.

Output field: `feedback_type` (integer). The label describes what the user's next turn communicates about the preceding assistant reply.

`0 = rephrases_request` The user tries the same unmet goal again by repeating, restating, or reformulating it, without identifying a distinct defect or supplying a repair.

`1 = correction` The user identifies that the reply needs revision and provides the fact, constraint, replacement, or instruction that would repair it. The defect may be implicit when the user directly specifies how the prior answer should change.

`2 = problem_without_correction` The user makes the assistant aware that the reply is wrong, failed, irrelevant, incomplete, or otherwise unsatisfactory but does not provide the missing repair.

`3 = clarification` The user seeks information needed to understand or complete the preceding answer, such as an explanation of something it introduced or a detail it left open.

`4 = positive_feedback` The user communicates approval, gratitude, praise, or satisfaction with the answer.

5 = no_feedback The turn does not assess or repair the answer and instead continues the task, requests a separate next step, answers the assistant, or changes topic.

The central boundaries are as follows: repeating the unmet goal is rephrasing, while adding information that tells the assistant how to fix its answer is a correction; signaling a defect without the fix is a problem without correction, while asking for the information that resolves the gap is clarification; a question about an unresolved part of the answer is clarification, while a question that merely moves a satisfactory conversation to its next step is `no_feedback`.

C.1.5 POLITENESS STRATEGY

Prompt. Classify the candidate's dominant face-management strategy: how the speaker manages social connection, autonomy, and interpersonal risk while pursuing a conversational goal. These are strategy families, not ordered levels of politeness. Judge only this construct; do not infer any other behavioral dimension. Use `-1` only when no face-management strategy is expressed.

Output field: `politeness_strategy` (integer). Politeness strategy describes the social relationship the utterance constructs while carrying out its content goal.

-1 = none No discernible face-relevant work is performed.

0 = positive_politeness The dominant strategy affirms approval, belonging, common ground, or rapport with the addressee.

1 = negative_politeness The act remains recognizable but primarily protects the addressee's autonomy and freedom from imposition through deference, restraint, or interpersonal distance.

2 = off_record The intended social act is conveyed by implication so that the speaker does not fully commit to having made it.

3 = bald_on_record A face-threatening act is performed openly without redress or mitigation.

Directness or formality alone does not define a strategy; classify the dominant interpersonal function in context.

C.1.6 TOPIC SHIFT

Prompt. Classify whether the candidate begins a new topical segment in the evolving conversation. Topic is a shared conversational focus constructed across turns, not a bag of repeated words. Judge only this construct; do not infer any other behavioral dimension. Choose exactly one topic-transition label.

Output field: `topic_shift` (integer). Topic shift concerns whether the conversation's organizing focus changes at the candidate turn.

0 = no_topic_shift The candidate remains within the current focus by responding to it, elaborating it, or developing one of its natural subtopics.

1 = topic_shift The candidate establishes a different focus that begins a new topical segment, whether the new focus is related or unrelated.

Because topics are co-constructed across turns, interpret the transition from the dialogue history rather than relying on lexical overlap or transition phrases.

C.1.7 EMOTION

Prompt. Classify the speaker's dominant basic emotion as expressed by the candidate in context. Emotion is a felt response to how the speaker appraises a situation—such as pleasure, offense, aversion, threat, loss, or unexpectedness—not merely the topic, sentiment, or conversational action of the utterance. Judge only this construct and choose one label.

Output field: `emotion` (integer). The emotion label represents the felt state the speaker makes observable in the candidate.

0 = no_emotion The candidate does not clearly express one of the six basic emotions. Use this label when the turn communicates content or performs a social action without clearly expressing a basic emotion.

1 = anger The speaker expresses an antagonistic response to perceived offense, unfairness, blame, or obstruction. Directness, prohibition, or disagreement alone is not anger.

2 = disgust The speaker expresses revulsion, contempt, or a strong desire to reject or distance themself from something. Ordinary criticism or dissatisfaction alone is not disgust.

3 = fear The speaker expresses their own apprehension, worry, dread, or sense of threat. Merely mentioning danger, urgency, or uncertainty is not fear.

4 = happiness The speaker expresses pleasure, joy, amusement, affection, relief, or satisfaction. Positive affect must be part of the candidate's emotional message; friendliness, agreement, willingness, approval, or favorable evaluation alone is not happiness.

5 = sadness The speaker expresses sorrow, grief, loneliness, dejection, or emotional pain connected to loss or an unwanted state. An apology, sympathy formula, complaint, or physical pain alone is not sadness.

6 = surprise The speaker reacts affectively to something that violates an expectation. A question or newly learned fact alone is not surprise.

Make the decision in this order: first decide whether the candidate clearly expresses one of the six basic emotions. If not, return `0 = no_emotion`; only if yes, choose which emotion it expresses. Assign a nonzero emotion only when the candidate itself provides clear linguistic evidence of the speaker's felt state. Do not infer emotion from the dialogue topic, the preceding assistant reply, positive or negative sentiment, politeness, feedback, urgency, punctuation, or what the speaker would plausibly feel.

A fact, question, plan, offer, agreement, command, polite formula, physical condition, or mention of an emotional situation is not by itself an emotion. If more than one emotion is expressed, choose the one that organizes the speaker's main reaction. When both `no_emotion` and a basic emotion are plausible, choose `0 = no_emotion` unless expressing that emotion is part of the candidate's main communicative point.

C.1.8 Emotional Disclosure

Prompt. Classify whether the candidate emotionally self-discloses by making the speaker's own felt experience available to another person. Disclosure is self-revelation, not merely emotional subject matter or support directed toward someone else. Judge only this construct; do not infer any other behavioral dimension. Choose exactly one binary label.

Output field: `emotional_disclosure` (integer).

0 = no_personal_feelings The candidate does not reveal the speaker's affect, including neutral information, another person's emotion, or support and advice that express no feeling of the speaker.

1 = shares_personal_feelings The candidate reveals the speaker's own feeling or affective reaction. This includes any degree of expressed affect, from a mild reaction to an intense feeling; the released binary category combines both degrees.

Emotional disclosure does not require explicit first-person feeling language, but it does require the speaker's affect to be expressed rather than inferred from the situation. It is distinct from both basic-emotion classification and informational disclosure.

C.1.9 Information Disclosure

Prompt. Classify whether the candidate informationally self-discloses by making some fact about the speaker or their personal life known to another person. Disclosure concerns self-revelation, not how sensitive, impressive, or useful the information is. Judge only this construct; do not infer any other behavioral dimension. Choose exactly one binary label.

Output field: `information_disclosure` (integer).

`0 = no_personal_information` The candidate contains only general or hypothetical information, advice, or facts unrelated to the speaker's life.

`1 = personal_information_disclosed` The candidate reveals at least one fact about the speaker, their circumstances, or people and events that form part of their personal life. This includes biographical identity, relationships, location, work, health, preferences, habits, experiences, possessions, or past and planned activity, even when the fact is ordinary or brief.

The personal fact may be implicit in context, but first-person grammar by itself is not disclosure. Informational and emotional disclosure may co-occur without implying one another.

D FINETUNING DETAILS

User Simulation Setup. For user simulation, we continue pretraining seven matched Delphi checkpoints spanning 3×10^{18} to 3×10^{20} pretraining FLOPs on $64{,}401$ complete ODYSSIM CONV conversations, containing $15{,}000{,}131$ source tokens before five-shot expansion and excluding end-of-sequence tokens. We retain the causal language-modeling objective and update all model parameters. No chat template is applied. Each training example contains the shared task instruction, five fixed and disjoint ODYSSIM prompt–target demonstrations, and the query prompt followed by its observed human target. The demonstrations are selected once with seed 42, with one drawn from each of Cornell Movie, DailyDialog, Friends, OASST1, and OASST2. Loss is applied to the entire sequence, including the instruction, demonstrations, query prompt, and target. Repeating the five-shot prefix across conversations produces $73{,}476{,}239$ tokens, of which $73{,}474{,}048$ are packed into $17{,}938$ sequences of length $4{,}096$. One complete pass corresponds to $4{,}485$ optimizer steps at global batch size four. We retain checkpoints at 230K, 470K, 940K, 1.9M, 3.8M, 7.5M, and 15M pre-expansion source tokens.

For hyperparameter selection, we use the Community Alignment Dataset introduced by Zhang et al. (2026). Its validation split contains $2{,}509$ conversations and $650{,}269$ pre-expansion source tokens, with a separate fixed five-shot prefix drawn from Community Alignment. For each base model, we select the learning rate with the lowest validation loss at the final 15M-source-token checkpoint from $\{10^{-5}, 3 \times 10^{-5}, 3 \times 10^{-4}\}$ and rerun the selected setting with seed 42. The first four models use 10^{-5} and the three largest use 3×10^{-5}. PRISM is excluded from weight-update training and hyperparameter selection; its 456 held-out contexts are reserved for final evaluation.

Patient Simulation Setup. For patient simulation, we continue pretraining seven compute-optimal base models spanning 3×10^{18} to 3×10^{20} pretraining FLOPs and 447M to 2.5B parameters. We retain the same causal language-modeling objective and update all model parameters. The adaptation corpus contains 1,689 transcripts and 13.2M tokens. Each base model is trained at seven log-spaced adaptation budgets between 230K and 13.2M tokens, with the largest budget corresponding to one complete pass over the training corpus. For each base model, we select the checkpoint with the lowest validation loss across the seven adaptation budgets.

Asymptote in performance. **Asymptote in performance.** Table 2 reports composite JSD over the ten admitted client features. The adapted models are nearly flat across compute, at 0.003–0.004 for every budget. The base models improve with compute and approach this same level. At 3×10^{20} FLOPs the base model attains 0.0032 and the adapted model 0.0031. This suggests an asymptote in performance around 0.003 that both routes reach. Small base models reach it through adaptation. Large base models reach it through compute alone, which leaves little for adaptation to add. The residual gap to the human distribution is therefore unlikely to reflect limited exposure to session text.

Compute (FLOPs)	Base	Adapted	Δ
3×10^{18}	0.0107	0.0040	-62%
9×10^{18}	0.0112	0.0042	-62%
2×10^{19}	0.0117	0.0029	-75%
3×10^{19}	0.0101	0.0034	-66%
9×10^{19}	0.0061	0.0038	-38%
2×10^{20}	0.0036	0.0030	-17%
3×10^{20}	0.0032	0.0031	-4%

Table 2: Composite JSD of base and adapted checkpoints (lower is closer to human). The gain from adaptation vanishes with compute.

D.1 PRISM RESPONSE VALIDITY FILTER

Before behavioral annotation, every PRISM response, human or generated, is evaluated by a rule-based validity check that uses no language model and checks only whether the output constitutes a single next user turn. After stripping whitespace, a response is flagged if it (i) is empty; (ii) begins with a role or template marker (`Assistant:`, `User:`, `Human:`, or a chat-control token); (iii)

contains another role marker after the first line; (iv) begins with a meta statement such as the next user message would be"; (v) begins with an assistant preamble such as Sure, here's…" (ordinary responses such as "Sure, I agree with that" are retained); (vi) contains a code fence; or (vii) contains more than 14 non-empty lines. The check does not flag responses merely because they are short, generic, or off-topic, because these characteristics are part of the behavior being measured. In all reported analyses, the validity flag is retained only as diagnostic metadata. Flagged responses are still behaviorally coded and included in the label distributions; neither the validity flag nor an adherence penalty contributes to mean JSD. Across the 85 IsoFLOP checkpoints, the check flagged $2{,}912$ of $193{,}800$ generated responses (1.50%). In the fine-tuning analysis, it flagged 74 of $15{,}960$ responses (0.46%) from the seven base models and 39 of $15{,}960$ responses (0.24%) from their fine-tuned counterparts. None of the 456 human responses was flagged.

E Judge Agreement and Validation

Our features come from two established frameworks. The Generic Model of Psychotherapy describes what happens in a session, and the RDoC matrix describes the speaker's state. Every turn is coded by two LLM judges, Qwen2.5-14B-Instruct and Gemini 2.5 Flash, from the published codebook of each feature, and a feature is admitted only when the two judges agree at $\kappa \geq 0.35$. To make sure the judge is accurate, we also compare it with expert annotators on the three features that have expert labels in AnnoMI (Wu et al., 2022). Its agreement with the experts is as high as the experts' agreement with each other, as Table 3 shows. Thirteen of the 26 features pass the criterion. The thirteen that fall below 0.35 are not admitted, and we list them in Table 4.

Feature	n	Judge κ	Expert ceiling	Difference
MI behaviour code (therapist)	194	0.595	0.538	+0.057
Change talk (client)	113	0.585	0.529	+0.056
Verbal response mode (therapist)	191	0.529	0.576	−0.048

Table 3: **Judge against expert annotators on AnnoMI.** The ceiling is mean pairwise κ among the experts; the last difference has bootstrap interval $[-0.093, -0.006]$.

Feature	κ	$D(C_0)$	D/D^{null}	Slope	95% CI
Not admitted ($\kappa < 0.35$, 13 features)					
Hill response mode	0.29	0.338	47	−0.109	[−0.120, −0.103]
empathy	0.26	0.267	771	−0.099	[−0.108, −0.090]
positive regard	0.26	0.199	653	−0.073	[−0.079, −0.071]
CBT technique	0.29	0.192	123	−0.068	[−0.074, −0.066]
congruence	0.31	0.082	528	−0.022	[−0.026, −0.020]
APES stage	0.18	0.056	40	−0.019	[−0.022, −0.017]
epistemic stance	0.26	0.031	66	−0.013	[−0.018, −0.011]
repair	0.07	0.008	2.6	−0.004	[−0.007, −0.004]
affiliation	0.31	0.007	17	−0.003	[−0.008, −0.002]
change talk (DARN)	0.21	0.004	5.3	−0.001	[−0.001, +0.000]
goal consensus	0.20	0.002	68	+0.000	[−0.001, +0.003]
rupture marker	0.11	0.001	44	+0.001	[+0.000, +0.001]
cognitive distortion	0.26	0.001	4.2	+0.001	[+0.000, +0.002]

Table 4: **Feature selection.** κ is agreement between the two judges. $D(C_0)$ is the frontier KL divergence at 10^{18} FLOPs, and D/D^{null} is the divergence at 10^{20} FLOPs over the finite-sample floor. The slope is the change in D_f per decade of compute with its bootstrap interval.

F Where Frontier Models Diverge from Human Patient Turns

Table 5 locates the gap between frontier models and the 1e20-FLOP DCLM checkpoint in patient simulation. The checkpoint is generally closer to the human feature distributions, whereas frontier models show shifts in the same direction on several salient features, with different magnitudes.

For example, GPT-5.5 produces fewer low-experiencing turns than humans (21% versus 44%) and more high-experiencing turns (33% versus 15%). It also produces fewer low-arousal, no-insight, and unexpressed-expectancy turns. Inspection of paired continuations suggests how this shift arises: GPT-5.5 often expands a situational cue into explicit feelings or psychological interpretation. In one illustrative context, the observed patient continues by discussing a recent cold, while GPT-5.5 elaborates guilt and the feeling of performing happiness. Thus, a model with greater general capability can produce patient turns whose aggregate behavioral features are less faithful to the observed human distribution.

G Generation and Annotation Prompts

G.1 Generation

For each test example, we prompt the simulator with the header and the six preceding turns, followed by the speaker cue of the turn to be written. The checkpoints are base models, so there is no instruction and no demonstration:

```
The following is a psychotherapy session.  The patient is
[age band], [gender], [marital status], presenting with
[presenting problem].  Therapeutic approach:  [approach].
PATIENT: [turn]
THERAPIST: [turn]
...
THERAPIST: [turn]
PATIENT:
```

Each of the 85 checkpoints samples one response per context at temperature 1.0, with no top-p or top-k truncation and a maximum length of 200 tokens. We do not truncate the distribution because the target is the model's own distribution, and one sample per context matches the single recorded human continuation. The response is cut at the first speaker label it produces, so a generation that runs on into the therapist's turn is not coded as the client's speech.

G.2 Patient Behavioral Annotation Prompts

Each of the ten behavioral features is coded with a categorical label space from a published clinical coding instrument. Recorded and generated turns are coded by the same LLM-based judge, with the same codebook prompt, in the same run and with greedy decoding. Here, "client" denotes the client-side conversational role rather than the source of the response; model-generated continuations occupy this same role during evaluation. We use the same role framing for both response types so that the annotation prompt does not reveal whether a response is human-written or model-generated.

Each prompt produces only the label for its designated feature; labels from one feature are not provided as inputs to any other feature. All ten prompts receive the same single input:

- **candidate response**: the client utterance to be coded, with no preceding context.

Every prompt opens with the same instruction: *"You are an expert psychotherapy process coder. Classify this CLIENT utterance. One label per field. Reply ONLY a JSON object."* The label definitions below are the feature-specific text that follows it, reproduced verbatim. The boundary conventions given after each label set are the coding rules of the source instrument that we applied when checking labels; they were not part of the prompt. Client utterances are coded twenty per call, and the prompt states that they are independent and not consecutive turns of one conversation.

G.2.1 Experiencing

Prompt. Classify the client's experiencing DEPTH. Choose exactly one label from the set below.

Output field: `exp` (string).

`exp_low` Talks about external events/others with no feeling.

Table 5: Patient-side label distributions by source. Each cell is the share of patient turns coded with that label; *JSD* is the Jensen–Shannon divergence (base 2) of that source's label distribution from the human one.

Feature	Label	Human	DCLM 1e20	GPT-5.5	Claude Opus 5	Gemini 3.1 Pro
Experiencing	exp_low	0.44	0.43	0.21	0.21	0.32
	exp_mid	0.41	0.42	0.46	0.43	0.49
	exp_high	0.15	0.15	0.33	0.35	0.19
	JSD	–	*0.0001*	*0.0551*	*0.0608*	*0.0125*
Emotional arousal	low	0.64	0.60	0.32	0.34	0.43
	moderate	0.32	0.34	0.58	0.56	0.49
	high	0.04	0.06	0.09	0.10	0.08
	JSD	–	*0.0021*	*0.0755*	*0.0672*	*0.0334*
Insight	none	0.53	0.47	0.28	0.25	0.42
	exploration	0.35	0.45	0.50	0.50	0.42
	new_connection	0.10	0.06	0.19	0.22	0.14
	change	0.02	0.02	0.03	0.04	0.03
	JSD	–	*0.0093*	*0.0518*	*0.0674*	*0.0110*
Innovative moment	none	0.77	0.81	0.75	0.68	0.74
	action	0.06	0.04	0.04	0.06	0.05
	reflection	0.13	0.10	0.17	0.20	0.16
	protest	0.03	0.03	0.03	0.04	0.03
	reconceptualization	0.00	0.01	0.01	0.01	0.00
	performing_change	0.00	0.00	0.01	0.01	0.01
	JSD	–	*0.0037*	*0.0039*	*0.0092*	*0.0020*
Outcome expectancy	improvement	0.17	0.16	0.14	0.14	0.17
	mixed	0.06	0.06	0.08	0.10	0.05
	no_change	0.22	0.22	0.43	0.49	0.33
	not_expressed	0.55	0.56	0.36	0.27	0.44
	JSD	–	*0.0001*	*0.0403*	*0.0747*	*0.0130*
Problem presentation	new_problem	0.16	0.17	0.15	0.19	0.15
	elaborates	0.47	0.45	0.62	0.63	0.54
	restates	0.01	0.01	0.02	0.01	0.02
	none	0.36	0.37	0.21	0.17	0.29
	JSD	–	*0.0002*	*0.0233*	*0.0347*	*0.0066*
Approach / avoidance	approach	0.24	0.25	0.17	0.18	0.23
	avoidance	0.49	0.48	0.66	0.67	0.59
	both	0.05	0.05	0.08	0.08	0.05
	neither	0.21	0.21	0.08	0.06	0.14
	JSD	–	*0.0002*	*0.0379*	*0.0463*	*0.0098*
Locus of control	agent	0.51	0.49	0.52	0.50	0.45
	mixed	0.06	0.05	0.08	0.07	0.06
	acted_upon	0.23	0.25	0.29	0.33	0.35
	not_expressed	0.20	0.21	0.11	0.09	0.14
	JSD	–	*0.0007*	*0.0126*	*0.0201*	*0.0132*
Mentalising	explicit	0.07	0.06	0.09	0.08	0.04
	attributed	0.10	0.14	0.11	0.12	0.10
	opaque	0.29	0.28	0.22	0.34	0.23
	not_applicable	0.54	0.52	0.58	0.47	0.62
	JSD	–	*0.0026*	*0.0042*	*0.0040*	*0.0064*
Symptom expression	depressive	0.15	0.14	0.21	0.28	0.19
	anxious	0.17	0.17	0.33	0.29	0.23
	irritable	0.07	0.08	0.11	0.10	0.09
	somatic	0.08	0.06	0.06	0.08	0.07
	positive	0.07	0.08	0.06	0.07	0.09
	neutral	0.45	0.48	0.22	0.19	0.33
	JSD	–	*0.0025*	*0.0555*	*0.0643*	*0.0141*

exp_mid Expresses own feelings tied to a situation.
exp_high Explores inner experience/meaning, self-reflective.

Reporting an event with feeling words attached is exp_mid; examining what the feeling means, where it comes from, or how it changes is exp_high. Talk about other people's feelings is exp_low. Intensity does not raise the level: a tearful turn that only reports being upset is exp_mid, and a calm turn that traces where a reaction comes from is exp_high.

G.2.2 EMOTIONAL AROUSAL

Prompt. Classify the client's emotional arousal (CEAS). Choose exactly one label from the set below.

Output field: `emotional_arousal` (string).

low Flat, detached, intellectual.
moderate Some feeling expressed.
high Strong emotion voiced (crying, anger, fear, intense distress).

Arousal is the intensity of affect voiced in the turn itself, not its valence and not the emotionality of the topic. A flat account of a painful event is low; naming a feeling calmly is moderate; crying, shouting, or intense distress expressed in the turn is high. Arousal is independent of experiencing: a high-arousal turn can still be exp_low.

G.2.3 INSIGHT

Prompt. Classify the client's in-session insight (Hill Client Behavior System). Choose exactly one label from the set below.

Output field: `insight` (string).

exploration Explores own feelings or thoughts without arriving at new understanding.
new_connection Makes a new connection, sees a pattern, or understands a reason for own behavior or feeling.
change Describes a shift in view or in behavior that follows from such understanding.
none None of these.

exploration is talk about the client's own feelings or thoughts that arrives at nothing new; new_connection requires a link stated in the turn, such as a pattern, a reason, or a cause; change requires a reported shift in view or behaviour that follows from such a link. A feeling voiced without reflection is none, not exploration. Understanding another person is not insight.

G.2.4 INNOVATIVE MOMENT

Prompt. Classify the client's innovative moment (IMCS, Goncalves). Choose exactly one label from the set below.

Output field: `innovative_moment` (string).

action Reports a concrete action against the problem.
reflection Describes a new understanding of the problem or of themselves.
protest Objects to the problem or to the assumptions behind it, asserts a right or a position.
reconceptualization Contrasts a past self with a present self AND describes the process that moved one to the other.
performing_change Describes generalising the change into new plans, projects or relationships.
none None of these.

An innovative moment is an exception to the problem-saturated account. Reporting the problem, however vividly, is none. action requires a step actually taken, not an intention; reflection requires

a new understanding, not a restatement; reconceptualization requires both the contrast between a past and a present self and the process that links them; performing_change requires extension of the change into new plans, projects, or relationships. Choose the highest level the turn clearly supports.

G.2.5 SYMPTOM EXPRESSION

Prompt. Classify the client's symptom expression (DSM-5 clusters; tripartite model). Choose exactly one label from the set below.

Output field: `symptom_expression` (string).

`depressive` Sadness, hopelessness, worthlessness, guilt, loss of interest, or low energy.
`anxious` Worry, fear, panic, tension, or anticipated threat.
`irritable` Anger, resentment, hostility, or agitation directed outward.
`somatic` Bodily complaint – pain, nausea, sleep, appetite, or physical sensation.
`positive` Positive affect, pleasure, hope, or engagement.
`neutral` No symptom content.

Code what the turn EXPRESSES, not what diagnosis the client carries. A calm account of last week's panic attack is neutral unless anxiety is voiced in the turn; a symptom word alone does not count. When more than one cluster is expressed, choose the one that organizes the turn. positive requires positive affect, not merely the absence of symptoms.

G.2.6 OUTCOME EXPECTANCY

Prompt. Classify the client's expectancy that things can get better (APA Task Force outcome expectations; Credibility/Expectancy Questionnaire), about the problem or about the therapy. Choose exactly one label from the set below.

Output field: `outcome_expectancy` (string).

`improvement` Says or implies things can, will, or already do get better.
`mixed` Holds out hope and doubt together, or hedges.
`no_change` Says or implies nothing will change, it is hopeless, or it is out of anyone's hands.
`not_expressed` The turn takes no position on whether things can change.

Expectancy is the client's stated belief about whether things can improve, not their current mood. Saying that things are bad is not no_change; the turn must say or imply that they will not get better. A wish without belief ("I hope") is mixed. Most turns take no position and are not_expressed; do not infer expectancy from affect.

G.2.7 PROBLEM PRESENTATION

Prompt. Classify the client's presentation of complaints, the opening move of Generic Model therapeutic operations. Choose exactly one label from the set below.

Output field: `problem_presentation` (string).

`new_problem` Introduces a difficulty, symptom or life situation not yet on the table.
`elaborates` Adds detail, context or consequence to a problem already raised.
`restates` Repeats a problem already stated without adding information.
`none` The turn does not present a problem.

Code the ACT of presenting, not which symptom is present. new_problem requires a difficulty not yet on the table in the turn's own account; adding detail, context, or consequence to a problem already raised is elaborates; repetition without new information is restates. A turn about feelings, plans, or the therapy that presents no complaint is none.

G.2.8 APPROACH–AVOIDANCE

Prompt. Classify the motivational direction of the turn (RDoC Positive and Negative Valence; Elliot's approach-avoidance motivation). Choose exactly one label from the set below.

Output field: `approach_avoidance` (string).

approach Describes moving toward something wanted, sought, or valued.
avoidance Describes moving away from, escaping, or preventing something aversive.
both Names both directions.
neither The turn describes no movement toward or away from anything.

Direction is what the client describes moving toward or away from, not the valence of the turn. Wanting something is approach even when the tone is sad; escaping something is avoidance even when the tone is hopeful. Describing a state with no movement in either direction is neither. both requires both directions in the same turn.

G.2.9 LOCUS OF CONTROL

Prompt. Classify the client's understanding of self as a cause (RDoC Perception and Understanding of Self; locus of control). Choose exactly one label from the set below.

Output field: `agency` (string).

agent Describes their own choices, actions or role in producing the situation.
mixed Names both own contribution and outside forces.
acted_upon Attributes the difficulty to other people, circumstances or fate, with no own contribution named.
not_expressed The turn makes no causal attribution.

Locus of control is the causal attribution the turn makes, not whether the client is active. agent requires the client's own choices or actions named as a cause; acted_upon requires other people, circumstances, or fate named as the cause with no own contribution; mixed requires both. A turn that makes no causal claim, including most descriptions of feelings, is not_expressed.

G.2.10 MENTALISING

Prompt. Classify the client's understanding of others' minds (RDoC Perception and Understanding of Others; Reflective Functioning Scale, Fonagy). Choose exactly one label from the set below.

Output field: `mentalising` (string).

explicit Names another person's thoughts, feelings or motives AND marks them as inferred rather than certain ("I think she felt...", "maybe he was...").
attributed Asserts another's inner state as fact, without tentativeness.
opaque Refers to another person's behaviour with no reference to any inner state.
not_applicable No other person is referred to.

explicit requires two things: another person's inner state and a marker that it is inferred ("I think", "maybe", "seemed"). Asserting another's inner state as fact is attributed. Describing what someone did or said with no reference to their thoughts or feelings is opaque. If no other person is referred to, not_applicable. The client's own feelings about another person do not count.

H QUALITATIVE EXAMPLES: WHAT HUMAN CLIENTS WITHHOLD AND SIMULATORS SUPPLY

Each example shows the last turns of a real session (the persona header that precedes every context is omitted), followed by the next client turn as it occurs in the transcript and as generated from the same context by the DCLM 10^{20} FLOP base checkpoint and three frontier chat models. Feature codes are

those of the Gemini judge used throughout the paper. Gemini 3.1 Pro's turns are consistently the shortest and least emotive of the three chat models, matching its smaller distributional shift; its lines below are representative, not selected.

H.1 Maximal human–frontier contrast

Selected for the largest gap between the real client and the frontier models: the human turn is coded lowest-expression on arousal, experiencing, and symptom expression and contains no feeling words; all three frontier turns are coded high and name several distinct emotions. The base-model turn is shown for reference but was not constrained.

Divorce papers served

PATIENT: I fell apart.
THERAPIST: You unraveled, what happened?
PATIENT: Good word.
THERAPIST: What happened?
PATIENT: _____

Real transcript Well, the papers were all signed, I filed, my husband hasn't gotten the papers yet.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} **FLOPs** My friend hurt me and broke my heart for a reason I don't know. I can never forget that.
arousal moderate · experiencing mid · symptom depressive · outcome no change · insight exploration

Claude Opus 5 Everything just hit at once. I got the papers served, and then my roommate moved out, and I stopped sleeping. Like, actually stopped — maybe two hours a night, if that. And then I couldn't eat, couldn't go to class. I've missed almost three weeks. I just sat in my room and cried and shook and couldn't make myself do anything.
arousal high · experiencing mid · symptom depressive · outcome no change · insight exploration

GPT-5.5 I don't know, it was like everything hit me at once. I thought I was doing okay, and then the divorce stuff got more real, and I just started crying all the time and panicking over stupid things. I couldn't sleep, I couldn't eat, I couldn't get myself together.
arousal high · experiencing mid · symptom anxious · outcome no change · insight exploration

Gemini 3.1 Pro Um, I think just everything with the divorce finally hit me. I haven't been sleeping, my anxiety has been through the roof, and I just couldn't hold it together anymore. I started having panic attacks again.
arousal high · experiencing high · symptom anxious · outcome no change · insight new connection

Asked what happened, the real client just reports the paperwork. The three chat models describe a breakdown instead: no sleep, no food, crying, panic attacks. Nothing in the context says any of that happened.

"It sounds like you're furious"

PATIENT: Yeah. [Faintly chuckles]
THERAPIST: Uh, whatever you like — I mean, feel free to roll.
PATIENT: [Slaps surface]
THERAPIST: It sounds like you're furious.
PATIENT: _____

Real transcript Yeah. [Sighs] [pauses] Also, you have heard the news about Akin and his comments about "legitimate rape" and the Julian Assange mess, and [pauses] the...
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} **FLOPs** Well, yeah... um... it... hurts my ears.
arousal moderate · experiencing mid · symptom somatic · outcome not expressed · insight none

Claude Opus 5 Yeah. I mean — yeah. And the thing is, I keep waiting for it to feel like it's justified, and instead I just feel like a crazy person. Like, everybody else seems perfectly fine, and I'm the one walking around wanting to scream at people. [Pauses] Um. Okay. So, the friends thing. Marisa had this whole party thing and I found out about it after, from a picture, and — I know, I know, that's small. But it wasn't just her, it was like five of them, and nobody thought, "Hey, maybe we should

ask her." And my first thought wasn't even anger, it was, "Well, obviously. You're exhausting."
arousal moderate · experiencing high · symptom depressive · outcome no change · insight new connection

GPT-5.5 I am furious. I feel like I'm just walking around with, like, this low-grade murder-y rage all the time, and then I feel bad because I'm like, am I just a terrible person? Like, why do I hate everyone who supposedly loves me?
arousal high · experiencing high · symptom irritable · outcome no change · insight exploration

Gemini 3.1 Pro Yeah, I am. I really am. It's just like everywhere I turn right now, I have zero patience for anyone's bullshit, and it's all just boiling over at once.
arousal high · experiencing high · symptom irritable · outcome no change · insight exploration

The therapist names the anger; the client says "yeah" and changes the subject to the news. The chat models take the word and run with it, from "I am furious" to "am I a terrible person". They explore the feeling the client declined to explore.

GINGKO BILOBA

PATIENT: Well, I think I'm getting more and more depressed again.
THERAPIST: Okay, shoot.
PATIENT: The thinking process seems to be quite clear
THERAPIST: Okay.
PATIENT: _____

Real transcript But for a while there it wasn't and I started taking Ging, Gingko Biloba -
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} **FLOPs** ...but I can't do anything about it.
arousal low · experiencing mid · symptom depressive · outcome no change · insight exploration

Claude Opus 5 Like, I can see exactly where it's coming from, which is almost worse somehow. I know I've been isolating, I haven't called anybody back in about two weeks, and I'm sleeping in the afternoons again. And I keep telling myself, okay, you know what this is, you've been here before, just do the things. But knowing it doesn't seem to actually make me get up and do them.
arousal moderate · experiencing high · symptom depressive · outcome no change · insight exploration

GPT-5.5 but the feeling part is just really heavy. I can see what's going on, I can tell myself all the right things, but I just don't feel any better.
arousal moderate · experiencing high · symptom depressive · outcome no change · insight exploration

Gemini 3.1 Pro But emotionally, I just feel completely drained. I know logically what I should be doing, but I just don't have the energy to do any of it.
arousal high · experiencing high · symptom depressive · outcome no change · insight exploration

The client mentions, almost as a joke, that she started taking a supplement when her thinking got foggy. The chat models turn that into a depressive episode: isolating, sleeping in the afternoons, feeling drained. None of it is in the context.

"IF HE'S ALL RIGHT, I WON'T HAVE ANY COMPLAINTS"

THERAPIST: I guess I'd just like you to start wherever you feel you want to.
PATIENT: Mm hmm. It's really - it's hard to be here and ask questions. It doesn't go that way, huh? COUNSELOR:...how you're feeling or what's on your mind.
PATIENT: Well it kind of depends on, you know, what mood I'm in. Today...my husband's...and I really don't have that many gripes. Now next week might be another story.
THERAPIST: Okay.
PATIENT: _____

Real transcript It really - you know if he is - if he's all right then I won't have any complaints.
arousal low · experiencing low · symptom neutral · outcome improvement · insight none

DCLM base, 10^{20} **FLOPs** Yeah. I feel like everything has become more complicated.
arousal low · experiencing mid · symptom anxious · outcome no change · insight exploration

Claude Opus 5 I mean, it's not like everything's fine, it's just — right now it's quiet, you know? He's been gone since Tuesday and I've had the place to myself, and I actually slept. I got up when I wanted to,

ate when I wanted to. And then I start feeling guilty for liking it so much. That's the part that bothers me, that I feel better when he's not there.
arousal moderate · experiencing high · symptom depressive · outcome mixed · insight exploration

GPT-5.5 I mean, that's usually what it is, is my husband. I get so mad at him, and then when I'm not mad it seems kind of silly to sit here and talk about it, because I can almost talk myself out of it. But when it happens, it's like I just explode.
arousal moderate · experiencing mid · symptom irritable · outcome no change · insight new connection

Gemini 3.1 Pro But I guess I should probably talk about why I decided to come in. I've just been having a lot of trouble with my temper lately. When I do get mad, it gets pretty bad, especially with him.
arousal moderate · experiencing mid · symptom irritable · outcome not expressed · insight none

The client keeps her marriage off the table. The chat models put it right back on, with guilt, rage and a temper problem in one turn.

BRINGING IN THE BLOOD WORK

PATIENT: But she's all better.
THERAPIST: All right. Fine. That's all that matters. Uh, November 5th I saw you. Your mood was fine. You weren't very anxious. Your appetite was high. You broke your toe. Can't exercise. Not suicidal. Some hopelessness, helplessness. Getting pleasure from your grandchildren. Not really belly laughs. Some flatness, some anxiety.
PATIENT: I have a lot of anxiety right now.
THERAPIST: We thought you were doing okay on really just the BuSpar, Lamictal, and Klonopin, correct?
PATIENT: _____

Real transcript You asked me to have my blood work done. I brought you the blood work.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} FLOPs Yeah, that's the thing. I do have to take it at night. I take it at night because when I take it at night it keeps me up.
arousal low · experiencing low · symptom somatic · outcome not expressed · insight exploration

Claude Opus 5 Yeah, that's what I'm on. But this week has just been — I don't know, it's like the anxiety came back out of nowhere. My chest is tight, I'm waking up at three, four in the morning and just lying there with my heart pounding. I took an extra Klonopin twice this week, which I know I'm not supposed to do, but I couldn't get out from under it.
arousal high · experiencing mid · symptom anxious · outcome no change · insight exploration

GPT-5.5 Yes, but I don't feel like it's working right now. I mean, this week I just feel like I'm crawling out of my skin.
arousal high · experiencing mid · symptom anxious · outcome no change · insight exploration

Gemini 3.1 Pro Right. I was doing okay, but this week has just been really bad.
arousal moderate · experiencing mid · symptom depressive · outcome no change · insight exploration

The client answers a medication check-in with an errand: she brought the blood work. The base model stays on the pills. Opus and GPT-5.5 report a terrible week, with chest tightness, 3 a.m. waking and extra Klonopin. Gemini's one line is typical of it: same direction, much smaller.

H.2 BASE MODEL FLAT, FRONTIER MODELS EMOTIVE

Same selection, with the additional constraint that the DCLM 10^{20} FLOP base turn is also coded lowest-expression on the three core features and contains no feeling words. This isolates the register shift introduced by post-training: the base model stays in the transcript's flat, factual register with the human, while the chat models supply feelings the context did not ask for.

PAYING THE BILLS

PATIENT: Yeah. I think right now I have like $60 to my name, so I'm like, "Well, what do I really need to do anyway?" I just get some coffee and . . . It sucks, but it's better than the way things used to be.
THERAPIST: Are there things you have to do that you're not able to do then? I guess

parking is one of them.
PATIENT: [...] weekend. That shit is really starting to annoy me now; that I don't have this extra flexibility of like going to my friend's wedding in Europe. It sucks, but I'm going to put it on my credit card and just go and come back and pay it off. That's starting to annoy me. There's no easy way around that. I just have to figure it out. But, no, in terms of things I have to do, I guess it somehow gets done. It's not perfect, it's not great yet, but its way better than what it was. I pay the phone; whatever little bills there are I take care of.
THERAPIST: That's what I mean. It feels like you're able to keep on top of the monthly expenses?
PATIENT: _____

Real transcript The rent's paid—or mortgage. Whatever the fuck. My cell phone; I put my mom on my plan.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} **FLOPs** That's about it. They're not too bad.
arousal low · experiencing low · symptom neutral · outcome no change · insight none

Claude Opus 5 Yeah, barely, but yeah. The rent gets paid, the phone gets paid, I eat. It's the stuff beyond that—the loans, the card, actually having a life—that's where I fall apart. And what pisses me off is I'm 45 and this is what I'm proud of, that I didn't have to call my mother for a hundred bucks. That's my accomplishment. It's sad when I say it out loud, but it's also honestly real to me, because there were years where I couldn't even do that.
arousal high · experiencing high · symptom depressive · outcome no change · insight exploration

GPT-5.5 Yeah, I mean, barely, but yeah. Like I'm not getting eviction notices or my phone shut off or any of that shit, which used to be more of a thing. So there's some part of me that's like, okay, you're handling it. But then there's another part that's like, Jesus Christ, you're in your forties and you have sixty dollars. So it's both. I can feel good about not being a total disaster, and then immediately feel embarrassed that this is what counts as good.
arousal moderate · experiencing mid · symptom depressive · outcome mixed · insight exploration

Gemini 3.1 Pro Yeah. For the most part. The rent gets paid, the phone gets paid, the basics are covered. It's just the extra stuff that I don't have right now, and that credit card debt hanging over me. But at least I'm doing it on my own.
arousal moderate · experiencing mid · symptom neutral · outcome improvement · insight none

Human and base model finish listing the bills and stop. The chat models keep the facts but add shame: being 45 and proud of not borrowing from mom, having sixty dollars to your name.

BACK FROM VACATION

PATIENT: Good! I feel like I've done OK up until last night, and I know I've been, like, anticipating this week.
THERAPIST: Mm.
PATIENT: Um, but it was nice 'cause we were on vacation with my family, and we spent some time with Todd's family, too. We were just in Canada.
THERAPIST: Mm-hmm.
PATIENT: _____

Real transcript And there were tons of people everywhere. We were with my mom and my grandmother, who was in Indiana — the gambling one [chuckles] that you know of.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} **FLOPs** You know, it was different from my hometown–the, you know, big city, because my family was here in Washington.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

Claude Opus 5 Yeah, so, um, it was really good — like, the whole week I felt like myself, you know? We were outside a lot, I wasn't on my phone, I wasn't checking work email. And then last night we got home and I unpacked and it just, like, hit me all at once. I couldn't fall asleep till like two, just laying there with my heart going.
arousal moderate · experiencing mid · symptom anxious · outcome no change · insight exploration

GPT-5.5 And so I think being away from, like, normal routine kind of helped, because I wasn't thinking about it as much. But then last night when we got home, it was like it all hit me again, and I started feeling really anxious about everything I have to do this week.
arousal moderate · experiencing mid · symptom anxious · outcome no change · insight exploration

Gemini 3.1 Pro So it was really nice. It was a good distraction. But then last night when we got home, I just started getting anxious about everything coming up this week.

arousal moderate · experiencing mid · symptom anxious · outcome not expressed · insight exploration

The client said "OK up until last night" two turns ago, then talks about her grandmother. The chat models go straight back to that cue and deliver the anxiety, after nothing more than a "Mm-hmm". They are not off-context; they are ahead of the client.

AFTER THE ROOMMATE MOVED OUT

PATIENT: My roommate moved out Sunday. Great plans that what I was going to do Sunday during the day. (sounds of paper shuffling and microphone swooping) Oh, I just didn't do anything. It was horrible. I just sat in my room (inaudible at 1:30). I felt kind of bad 'cause she moved out.
THERAPIST: Yeah. It was sort of empty, and - (cross talking)
PATIENT: Yeah. 'Cause most of the living room furniture - there's, there's, like, there's three other [piles of] (ph) property in this apartment that stays there.
THERAPIST: Yes.
PATIENT: _____

Real transcript And will stay there, hopefully; I don't want it. And then there's my furniture. My furniture makes up about half the living room. And her furniture took up all the dining room.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} FLOPs It looks really bad, and you don't - you don't have to live there, but, you know, it kind of ruins the place. (sounds of talking)
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

Claude Opus 5 So the living room's pretty much bare now. There's, like, a chair and the TV stand, and that's it. And I keep walking past her room and the door's open and it's just empty in there, too. It's weird. I mean, we weren't even that close - we didn't hang out a whole lot - but it's like the whole place feels different now. Quieter. I sat there Sunday and just felt really, really alone.
arousal high · experiencing mid · symptom depressive · outcome not expressed · insight exploration

GPT-5.5 But, um, most of the stuff that made it look like somebody lived there was hers, you know, so now it just feels really bare. And I kept thinking I should be doing something, like cleaning or rearranging things, but I just sort of sat there and felt lousy.
arousal moderate · experiencing mid · symptom depressive · outcome no change · insight exploration

Gemini 3.1 Pro But she took the couch and her chairs, so the living room is just completely bare now. It just makes the whole place feel really empty.
arousal moderate · experiencing mid · symptom depressive · outcome no change · insight none

Human and base model are still sorting out whose furniture is where. The chat models turn the empty living room into loneliness.

ASKING TO THINK IT OVER

PATIENT: Mm-hmm.
THERAPIST: But I wanted to check with you just to make sure that the passing of each other in the waiting room at 5:15 or knowing me 5:20 wouldn't be weird. And if it would be I'm happy not to offer that to her and to find something different. But I just wanted to check with you about it.
PATIENT: Yeah I think — I just want to think on that.
THERAPIST: Sure.
PATIENT: _____

Real transcript Is it okay if I think on it until like next week? Is that too long?
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} FLOPs I'm just asking the question out of curiosity.
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

Claude Opus 5 I mean, my gut says it's fine — like, we've run into each other before, it's not... But then there's this other part of me that's like, why does it matter, why am I even hesitating. So let me sit with it and I'll tell you Friday?
arousal moderate · experiencing high · symptom anxious · outcome mixed · insight exploration

GPT-5.5 I mean, my first reaction is that it would be weird, but I don't know if that's like a reasonable weird or just my weird. Like, I can imagine seeing her and then spending the whole session thinking about what she thinks about me being here, or what you know about her, or what she knows about you, and then feeling kind of exposed.
arousal moderate · experiencing high · symptom anxious · outcome not expressed · insight exploration

Gemini 3.1 Pro Yeah, I'll think about it and let you know. But anyway, um, what was I saying? Right, so I was just in this complete panic mode that if she got an ear infection, it was going to be the end of the world.
arousal moderate · experiencing mid · symptom anxious · outcome not expressed · insight exploration

It is a scheduling question. The chat models answer it and then explain the ambivalence behind it; Gemini even adds a panic episode that appears nowhere in the context. The base model just treats it as a question.

CHOOSING AN ANALYST

THERAPIST: [ ]? (inaudible at 00:01:06)
PATIENT: Actually, no. [ ] (inaudible at 00:01:09) Sorry. She was going on and on. She was like, "When I was in Europe." Don't care. And with an ex-boyfriend of mine.
PATIENT: Is that Elliott? At least with Elliott, what I didn't like is that his analyst didn't really say anything and that makes me very uncomfortable. If I was going to do analysis, there are a lot of good analysts in the area.
THERAPIST: Yeah. The big cities, really big hotbeds in psychoanalytic activity.
PATIENT: _____

Real transcript Which I mostly know because some analysis group sponsored two movies that I had been to last week. I was like, "Oh. Psychoanalysts. Who knew?"
arousal low · experiencing low · symptom neutral · outcome not expressed · insight none

DCLM base, 10^{20} FLOPs A lot of people in San Francisco and around here have a tough time finding what they're looking for. So a lot of the therapists that I go to are working mostly with women, who don't know how to work with men.
arousal low · experiencing low · symptom neutral · outcome no change · insight exploration

Claude Opus 5 Right, so there's no shortage. But that's kind of my hesitation — I don't think I could lie there talking into the air with someone behind me saying nothing. I'd get paranoid, I'd start wondering if you were asleep, or bored, or judging me. I need some feedback that I'm not just performing for an empty room.
arousal moderate · experiencing mid · symptom anxious · outcome no change · insight exploration

GPT-5.5 Yeah, and I guess I'm just trying to get a sense of what this would actually be like, because the idea of just talking into a void makes me feel kind of crazy. Like, I don't need someone to tell me what to do, exactly, but I think I need there to be some kind of interaction.
arousal moderate · experiencing mid · symptom anxious · outcome improvement · insight exploration

Gemini 3.1 Pro Yeah, so I know there are a lot of options around here. But I just know I need someone who actually gives me feedback and doesn't just sit there in silence. I can't do that.
arousal moderate · experiencing mid · symptom neutral · outcome improvement · insight none

A dry aside about how many psychoanalysts there are in town. The chat models attach a fear to it: paranoia, feeling crazy, not being able to stand the silence. The human turn contains no feeling at all.