Toward a Foundation Model of Human Behavior

Jonathan Politzki
Jean Technologies • jonathan@jeantechnologies.com

Abstract. How might deep learning offer a modern answer to old questions about understanding the human mind? Trained on vast amounts of human text, large language models learned to model the people who wrote it. We explore three fields, human-computer interaction, simulation, and recommendation, that predict, simulate, and model humans for a wide range of tasks, and argue that they are converging on the same model: a *foundation model of human behavior*. Two of the three ingredients that carried deep learning elsewhere, a generic objective and scaling, are already in place; the third, a corpus of what people do that spans the silos holding it, is not. Our contributions are a checkable definition that rolls the fields' own into one, a survey of what each field has built, a formalization of their convergence into a single next-action model, and directions toward the corpus.

1 Introduction

1.1 Motivation: understanding human behavior

Psychology, cognitive science, sociology, anthropology, and economics have each sought to understand the human mind and the processes that produce our behavior. The goal has largely evaded them: behavior is high-dimensional and context-dependent, it is observed only through sparse and fragmented traces, and the verbal theories built atop those traces underdetermine prediction.

In recent decades, deep learning has shown a remarkable ability to solve problems across domains. Given large amounts of data and general methods of computation, deep learning matches or exceeds human ability at tasks that had resisted every hand-built method; the 2024 Nobel Prizes in Physics and Chemistry were both awarded for work in or enabled by neural networks. The Chemistry prize is the instructive case. AlphaFold solved a fifty-year-old problem by supervised training on solved structures (Jumper et al., 2021); the models that followed it are foundation models in the sense of §1.3: a language model trained only to fill in masked residues of protein sequences learns to predict structure from a single sequence (Lin et al., 2023). Understanding of the protein fell out of predicting its sequence.

In recent years there has been growing interest in applying these methods to psychology, cognitive modeling, and the prediction of human behavior. The appeal is practical: deep learning offers a way onto problems that humans cannot solve directly, or could not solve on any reasonable timescale. This paper examines whether the same holds for understanding people, with deep learning models, and specifically large language models, as the candidate.

Our contributions are fourfold:

1. to roll the definitions the fields have offered into one checkable definition of a foundation model of human behavior (§1.4, Table 3);
2. to survey three fields building toward it, human-computer interaction, simulation, and prediction and recommendation (§2);
3. to formalize their convergence: each field's formula reduced to one next-action model, with what each field keeps of the person's stream made explicit (§3);
4. to outline directions that complete the circle back to the goal: a computer that understands people, reached by unifying the data held in separate silos, in one language, into one model (§4).

1.2 What it means to "understand"

Understanding cannot be cleanly attributed to machines, and the discussion slides easily into epistemology (Searle, 1980; Bender and Koller, 2020); Bommasani et al. (2021, §2.6) structure this debate for foundation models specifically, distinguishing what understanding would mean in principle from how we could detect it in practice. For the purposes of this paper, a computer understands a person insofar as it can predict or simulate what that person will do in any situation. This is Newell's criterion for a unified theory of cognition: a single model that behaves as an (almost) arbitrary function of the environment, rather than a separate theory per task (Newell, 1990).

The model. Let h_u denote the recorded stream of a person u: the actions they took and the situations they took them in, as one sequence. We assume the stream is rich enough that the current situation is simply its most recent portion. Deployed systems hold a finite window over fragmented data, and long inputs degrade performance well before the window is full (Hong et al., 2025); memory systems (Packer et al., 2023; Chhikara et al., 2025) and context engineering bridge the gap, and they matter because the trained model fixes an upper bound on performance and supplying the right context is how a system reaches it. This paper is about the bound. The model is a single function, trained end to end, that predicts the next action from the stream,

$$p_\theta(a_{t+1} \mid h_u)\,, \tag{1}$$

where a_{t+1} is the next action. Computing Eq. 1 forces the stream through an internal summary: the network compresses

h_u into a state

$$z_u = f_\theta(h_u) \tag{2}$$

and predicts from that state. Nothing about z_u is designed or trained separately: it is whatever summary of the person the model learns because prediction demands one. The predictive accuracy of the model hinges on the quality of z_u: a good representation captures the underlying explanatory factors of the person's behavior, remains useful across many downstream tasks, and disentangles what is stable about the person from what is situational (Bengio et al., 2013).

Prediction and simulation. We use these terms for two uses of the same model. Prediction evaluates Eq. 1: given the stream, score what the person does next. Simulation samples from it, feeding each sampled action back into the stream to roll forward. When the simulated individual is a real person, simulation is graded as prediction, against held-out behavior (Park et al., 2024); simulation also runs where prediction cannot, populating environments with synthetic agents and playing out counterfactuals (Park et al., 2023).

Generality. In line with the definition above, a good z_u should carry predictive power for a wide range of tasks, including tasks absent from the training distribution. Formally,

$$G(z) = \mathbb{E}_T \left[\text{Perf}_T(z)\right], \tag{3}$$

the expected performance of the representation across downstream tasks T. G is defined only relative to the set of tasks: "more general" means nothing until that set is fixed, which is what a benchmark does (§4).

1.3 Foundation models

The term *foundation model* was coined in 2021 by Stanford researchers: "a foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks" (Bommasani et al., 2021). The definition has three checkable clauses: broad data, generic self-supervision, and adaptation to many tasks. The same report names the two properties that make such models significant: *emergence*, behavior implicitly induced rather than explicitly constructed, and *homogenization*, one model coming to underlie many applications. Four ingredients recur in how these models are built.

Self-supervision. The training signal is the data itself, so the method scales to whatever data exists.

Tokens and the transformer. The dominant architecture discretizes its input into tokens and models the sequence autoregressively with attention (Vaswani et al., 2017). A large language model is trained to predict the next token; the representations it learns are useful far beyond that task, which is generality (Eq. 3) in the sense of §1.2.

Scaling. Loss falls as a power law in model size, data, and compute (Kaplan et al., 2020). Training proceeds in stages: large-scale pre-training under the self-supervised objective, followed by post-training that aligns the model to use (Ouyang et al., 2022).

Data. Scaling makes data the binding ingredient: compute-optimal training grows the token count in proportion to model size (Hoffmann et al., 2022), the corpus is the public web, and projections have frontier models exhausting the stock of public human text within this decade (Villalobos et al., 2024). Selection and quality matter alongside volume (Albalak et al., 2024). For language, the corpus existed before the models did; every ingredient above is generic, so where a comparable corpus exists, a foundation model can follow.

1.4 A foundation model of human behavior

A **foundation model of human behavior** is a model that

- D1. is trained on *behavioral traces*: timestamped records of actions people actually took, in context;
- D2. under a *generic self-supervised objective*, not a task-specific supervised target;
- D3. such that training *learns a person-level representation* $z_u = f(h_u)$ from the person's history alone;
- D4. and the *one resulting artifact serves many tasks*: prediction, ranking, simulation, profiling, etc.

D2 and D4 carry over the clauses of §1.3; D1 and D3 say what the data and the product are. The definition is a roll-up: the fields have each offered their own, a foundation model of human cognition, a general user model, a large behavioral model, and this one keeps what they share and drops what is particular to one field. Those definitions, with the near-synonyms and false friends around them, are collected in Table 3 and Appendix A.1. Next-action prediction is the leading candidate objective in the literature, not part of the definition.

Does an LLM already qualify? Yes. Text is human-generated, so a model of text is a model of human behavior, though its training never saw a per-person stream. Kosinski (2024) draws the consequence: to predict the next word of a sentence generated by a human, a model "must model not only grammar and vocabulary but also the psychological processes humans use when generating language." The empirical record bears this out, from theory-of-mind performance (Kosinski, 2024), contested on robustness by Ullman (2023), to reproduced behavioral experiments (Horton, 2023; Hewitt et al., 2024).

Its limits as a model of people. There are limitations. These models were not built for the role:

- *Objective:* post-training collapses the model into a single helpful assistant, narrowing the human distribution pre-training absorbed (Ouyang et al., 2022; Santurkar et al., 2023; Padmakumar and He, 2024; Hagendorff et al., 2023);
- *Data:* pre-training sees documents with no person attached, never the per-person stream h_u of D3 (Lei et al.,

2026), while representing a person takes long per-person sequences (Jiang et al., 2026; Pi et al., 2019; Pancha et al., 2022);

- *Signal:* text records what people say, not what they do; the behavioral-modeling literature claims that actions carry signal words lack (Zhai et al., 2024; Brüel Gabrielsson, 2026a).

Whether these differences are fundamental, or an LLM alone already suffices, is taken up in §5.

2 Survey: three fields

With the definition of §1.4 in hand, we survey three fields in which models meeting it, in whole or in part, have been built:

- **Human-computer interaction** observes a single user deeply and maintains a standing model to assist them (§2.1);
- **Simulation** asks what a person or a population would do in a given situation. It is the most open-ended of the three: the situation may be an environment the simulated agents live in, and the aim is prediction in any situation, which is the generalization form of Eq. 1. It has a commercial category of its own (§2.2);
- **Prediction and recommendation** trains on billions of users' interaction sequences to predict the next action (§2.3).

Each section runs in the same order: how the field started and what it aims at, what has been built, how it is measured, and its limitations. What each field gains over a base language model is collected once, in §3.2.

2.1 Human-computer interaction (HCI)

Origin and aim. HCI studies how people use and are served by computing systems. Its user-modeling tradition is as old as interactive computing: the first user-modeling paper built stereotype models of individual users in order to recommend novels (Rich, 1979). The aim has not changed. A system that serves a person should maintain a model of that person, and the field's recent position paper states the motivation plainly: a gulf stands between what users intend and what they manage to specify, and one way to close it is a better model of the user (Ziems et al., 2026), a model that knows who we were, who we are, and who we want to become.

The models. Three forms are in use.

- *Assistant memories.* Commercial chat assistants keep background-curated user memories built from chat histories and interaction events, reviewable by the user; documented as products rather than papers, so noted without citation.
- *General user models.* GUM takes screenshots of any computer use in and produces confidence-weighted natural-language propositions about the user's preferences, context, and intent, held on device and consultable by any assistant (Shaikh et al., 2025). Nothing is trained.
- *Next-action predictors.* LongNAP trains (Shaikh et al., 2026): from a month of screen recordings per user, a vision-language model learns by reinforcement to retrieve from a per-user memory and predict the user's next actions, rewarded by judged similarity to what the user actually did.

Training. There is no behavioral pretraining stage in this field. GUM trains nothing. LongNAP starts from a vision-language model pretrained on text and images and post-trains it by reinforcement, with the judged similarity of predicted to actual actions as the reward (Shaikh et al., 2026). The behavior enters only at adaptation time.

The formula. HCI predicts the user's next action from a curated memory of the stream:

$$p_{\text{LLM}}(a_{t+1} \mid m_u,\, c_t)\,, \qquad m_u = g(h_u), \tag{4}$$

where m_u is the memory (GUM's propositions, LongNAP's per-user store) read by a pretrained model, and c_t the current screen. m_u is an engineered stand-in for the learned z_u of Eq. 2. Inputs are screenshots, but both convert observation to language before modeling; the field has not moved past language tokens.

How it is measured. Neither GUM nor LongNAP has a standing benchmark; each is graded against its own users. GUM is judged by the people it models: participants rated propositions built from their own email streams, with maximum-confidence propositions rated fully accurate, and a five-day deployment of its proactive assistant replicated the pattern on live screen observations (Shaikh et al., 2025). LongNAP holds out by time and measures LLM-judged similarity to the user's actual next actions plus human win rates (Shaikh et al., 2026). Shuffling the history's order costs accuracy, the sequential signal Eq. 4 assumes.

Limitations. The stream satisfies D1 for one person at a time, with consent and depth the other fields lack, and a representation the user can read and correct. The published models train on tens of users, and today's collections sit under research protocols that prevent pooling. Population-scale training on such streams has not happened yet; nothing about the method forbids it, and assembling that corpus is the open step (§4).

2.2 Simulation

Origin and aim. Simulating people is an old ambition; what is new is the agentic instantiation: an LLM given memory, reflection, and planning, placed in an environment with other such agents, produces believable individual and group behavior (Park et al., 2023), after earlier work prototyped social systems with simulated communities (Park et al., 2022). Economics arrived independently, re-running classic behavioral experiments on LLM subjects and checking the results against the

published human ones (Horton, 2023); conditioned on demographic backstories, sampled responses track subpopulation distributions (Argyle et al., 2023). The aim is agents whose behavior matches people's in any situation, and the standard in this early work was believability or aggregate fidelity: no particular person was being modeled.

The models. Two lines. The prompted line supplies the person in context: a demographic profile (Argyle et al., 2023), or a two-hour interview transcript, which is what the agents of Park et al. (2024) are conditioned on. The trained line moves the person into weights. HumanLM trains a simulator to align a latent user state (stance, emotion) with the person's actual responses rather than imitate surface text (Wu et al., 2026); HumanLLM fine-tunes on the longitudinal public traces of 282K accounts (Lei et al., 2026); OdysSim builds a foundation model for behavior simulation outright (Zhou et al., 2026). Training is displacing prompting.

Training. The trained line post-trains a language model. HumanLM fine-tunes on persona, situation, and response triples and then optimizes, by reinforcement, the alignment of a latent user state with the person's actual responses (Wu et al., 2026). HumanLLM fine-tunes on its per-account traces and then merges the result one-to-one with the original weights to limit forgetting of general capability (Lei et al., 2026). As in HCI, the behavior enters at adaptation time; the pretraining is on text.

The formula. The field's data takes the shape HumanLM makes explicit, triples of persona, situation, and response (Wu et al., 2026), a schema the interview agents fit without stating. The situation often includes an environment, sometimes a closed one with a fixed set of options, a survey item or a game move. The model is

$$p_{\text{LLM}}(y \mid s, \pi_u), \tag{5}$$

a response y to a situation s given a *supplied* persona π_u rather than a learned representation. Input and output alike are language tokens.

How it is measured. The yardsticks are the most person-grounded of the three fields. Park et al. (2024) interview 1,052 people and grade the agent by normalized accuracy: its accuracy on held-out survey responses divided by the accuracy with which the participants replicate their own answers two weeks later; interview-conditioned agents reach 0.83 of that ceiling. HumanLM ships a 26K-user benchmark scoring alignment against held-out responses (Wu et al., 2026); trained simulators are also scored on out-of-domain social benchmarks (Lei et al., 2026), and OdysSim ships a capability taxonomy (Zhou et al., 2026). Elicited data, interviews and experiment trials, is the substrate throughout; the trained line is beginning to add public traces.

Limitations. Three. Grounding: a described persona is not the person. Prompted models reflect the opinions of some demographic groups far more than others (Santurkar et al., 2023), their survey answers carry artifacts of their own (Dominguez-Olmedo et al., 2024), and the line separating the positive results from the nulls is conditioning on an individual's actual data. Sensitivity: simulators diverge from humans under small rewordings of the same task (Schröder et al., 2025). Confidence: validation practice across the field is thin (Larooij and Törnberg, 2026), and no simulator reports calibrated confidence in what it predicts.

2.3 Prediction and recommendation

Origin and aim. Recommendation predicts what a person will engage with next, from what they and others engaged with before. Its history is a sequence of four pivots, each answering a limit of the last.

Filtering. Content-based filtering, an outgrowth of information retrieval, describes items by their features and matches them to a profile of what a person has liked; retrieval is the query-present case of the same problem (Belkin and Croft, 1992). Collaborative filtering answered content's limit by using other people's reactions: people who agreed in the past tend to agree again (Goldberg et al., 1992; Resnick and Varian, 1997). It worked because the signal is cheap and everywhere, and Amazon's item-to-item version made the field industrial (Linden et al., 2003). The Netflix Prize made matrix factorization the standard method: a person as a learned vector, a rating as its product with an item vector (Koren et al., 2009), a form psychometrics had written decades earlier (Table 4).

Deep learning. Wide & Deep named the field's central tension and marked its turn toward generalization: a wide part memorizes specific co-occurrences, a deep part generalizes through learned embeddings, and the two are trained jointly (Cheng et al., 2016). YouTube reformulated recommendation as extreme multiclass classification over a deep network (Covington et al., 2016), and attention over the user's action history followed (Zhou et al., 2018).

Sequence models. The objective became autoregressive: predict the next item from the sequence so far, with recurrent networks (Hidasi et al., 2016) and then transformers (Kang and McAuley, 2018; Sun et al., 2019). The person's representation became the sequence model's state, and that state was pretrained and transferred: across tasks (Yuan et al., 2020), across a retailer's surfaces (Shin et al., 2021), and as one user embedding serving a whole platform (Pancha et al., 2022). The term *foundation model* was not yet in use; the practice was.

The generative turn. After the success of large language models, researchers sought to bring their general world knowledge into recommendation, and with it the whole recipe: attention, autoregressive prediction, pretraining, and post-training with rewards. P5 cast recommendation as text-to-text (Geng et al., 2022); TIGER replaced item IDs with semantic IDs, learned codes an autoregressive model can generate (Rajput et al., 2023); HSTU recovered power-law scaling on raw event streams (Zhai et al., 2024). Deployment followed: PLUM adapts a pretrained language model to semantic-ID retrieval at YouTube (He et al., 2025), OneRec replaces Kuaishou's cas-

cade with one end-to-end generative model (Kuaishou OneRec Team, 2025), later released as open checkpoints (Zhou et al., 2025), and LLM-backed rankers at Netflix and JD beat mature production baselines (Li et al., 2026; Zou et al., 2026).

Training. This is the one field that pretrains on behavior, and its pipeline mirrors the language model's. OneRec builds a tokenizer, pretrains by next-token prediction over semantic IDs on the event stream, roughly 18 billion samples a day, and then post-trains by on-policy reinforcement with preference, format, and industrial rewards (Kuaishou OneRec Team, 2025). PLUM starts from a pretrained language model, expands its vocabulary with semantic IDs, continues pretraining on a half-behavior, half-metadata corpus, and then fine-tunes for retrieval with a reward-weighted loss (He et al., 2025). OpenOneRec aligns item tokens with text, co-pretrains, and adds a distillation stage in post-training to restore general reasoning (Zhou et al., 2025). HSTU trains in one pass over the streaming log and reports no post-training at all (Zhai et al., 2024).

The formula. Recommendation predicts the next item from the interaction sequence itself:

$$p_\theta(i_{t+1} \mid i_{1:t}), \tag{6}$$

the field in which the representation is routinely learned from behavior at population scale: z_u is the sequence model's hidden state, exactly the $f_\theta(h_u)$ of Eq. 2. The prior is population co-consumption absorbed into θ; the generative line adds a text prior. The deployed models' exact objectives are inventoried in Table 4.

Tokens. Unlike the other two fields, recommendation chooses its token. Raw item IDs carry no content, so a new item starts from nothing. TIGER's semantic ID quantizes an item's content embedding into a short code, so similar items share a prefix and a new item has a code from the day it is listed (Rajput et al., 2023). PLUM's tokenizer adds a contrastive term over co-engagement, so the code reflects both what an item is and who else chose it (He et al., 2025). OpenOneRec interleaves such codes with natural language in one language-model context (Zhou et al., 2025). The arc runs from IDs a language model cannot read to tokens it can.

The tokenizer is where much of the downstream performance is decided, and the field says so. PLUM states that the efficacy of a generative retrieval model "is fundamentally dependent on the semantic richness and structural integrity" of the semantic IDs, and spends a whole pretraining stage grounding the new tokens so they are "aligned with existing text tokens of the base language model" (He et al., 2025); OneRec rebuilds its tokenizer around collaborative signal because codes from content alone reconstruct items poorly (Kuaishou OneRec Team, 2025). The codes also constrain what the model can express: because generation walks a tree of codes, items that share a prefix receive similar probabilities for any user, and the model cannot represent some preference patterns that plain collaborative filtering captures (Hou et al., 2026). Catalogs change daily, so the codebook has to be maintained against drift as well as built.

How it is measured. Offline, by Recall@K and NDCG@K on held-out interactions, with Amazon Reviews and MovieLens the academic defaults (Rajput et al., 2023); online, by A/B lifts in engagement, the standard that decides deployment (Zhai et al., 2024; Kuaishou OneRec Team, 2025). The gap between the two is a standing problem the field names as such (Bougie et al., 2026). Open logs at billion-event scale now exist (Yandex, 2025; Ni et al., 2025; Yuan et al., 2022), and two instruments approach a standing benchmark: NineRec for transfer across domains (Zhang et al., 2023) and RecIF-Bench, eight tasks from prediction to reasoning released with OpenOneRec (Zhou et al., 2025).

Computation and efficiency. Recommendation serves far more queries per second than a chat assistant and trains on more events than there are words on the web, so it is engineered for cost, and the generative turn was partly an efficiency gain. Aggregating the deployment reports: model-FLOPs utilization rose from single digits under the classical cascade (4.6% for Kuaishou's ranking stack, 3.3% for its search stack) to the mid-twenties end to end, against roughly 40% for language-model training; operating expense fell to a tenth for recommendation and by three quarters for search; parameters moved out of embedding tables and into the network, 0.4% of parameters in the network for YouTube's embedding model against 90% for PLUM; and the sequence encoder runs 5 to 15 times faster than a standard transformer at long sequence length, which is what let Meta serve a 285-times more complex ranking model within a constant inference budget (Kuaishou OneRec Team, 2025; Chen et al., 2025; He et al., 2025; Zhai et al., 2024). Behavioral models are cheaper per query than language models and, so far, less efficient per FLOP.

The trajectory. Retrieval and filtering were argued to be one problem in 1992 (Belkin and Croft, 1992). In 2025 the company behind OneRec deployed the same generative recipe for search (Chen et al., 2025), and Netflix's ranker reads the user as verbalized context: the problem shifts "from feature engineering to context engineering," and "innovation moves up a level: from per-task architecture design to questions of data, scaling laws, post-training strategy, and inference optimization" (Li et al., 2026). Search conditions the person-model on an explicit query, recommendation on none, and chat elicits the query interactively; the three are converging into one autoregressive model consulted through different interfaces, one that can be steered in natural language and that generalizes to a cold start. The convergence runs one way so far: recommendation absorbed the language models' playbook, and nothing recognizable has flowed back. §4 takes up the reverse.

Limitations. D1 to D3 hold at population scale, with z_u learned from behavior alone and scaling laws to show for it (§3.3). Generalization has been demonstrated within one platform's surfaces and, through content tokens, across catalogs (Zhang et al., 2023; Zhou et al., 2025); it has not been demonstrated beyond

that, which is not the same as being impossible. The field's own description of itself is the point to hold: constrained by isolated data, today's recommenders "operate as domain specialists" (Zhou et al., 2025). §5 returns to it.

3 A foundation model of human behavior

Each field's models satisfy the definition of §1.4. HCI's next-action predictors train on a person's screen stream (D1), by next-action prediction (D2), into a per-user memory and weights (D3), serving prediction and assistance (D4). Trained simulators train on elicited responses and public traces, by predicting the response, into weights, serving prediction and simulation. Recommenders train on interaction logs, by next-item prediction, into a hidden state, serving retrieval, ranking, and cold-start scoring; the field already calls these models foundation models, with or without a language-model backbone underneath (Netflix Technology Blog, 2025; Zhou et al., 2025). One difference the training paragraphs of §2 expose: only recommendation pretrains on behavior. HCI and simulation post-train a language model on it, so strictly, in those two fields the foundation model is the language model and the behavioral stage is adaptation. Recommendation has built a foundation model from behavior; the other two have adapted one to it. The question that remains is whether one model underlies the three, and the formulas say it does.

3.1 One model, three settings

The paper's model is Eq. 1, restated:

$$p_\theta(a_{t+1} \mid h_u)\,,$$

predict what the person does next from their stream. Each field's formula collapses into it once the stream is written out.

The stream, written out. The stream h_u of §1.2 has four parts,

$$h_u = (\pi_u,\ E,\ s_{1:t+1},\ a_{1:t})\,: \tag{7}$$

the *persona* π_u, what is known about the person before the record begins; the *environment* E, the setting and its rules, a catalog, an app, a survey instrument, a game world; the *situations* $s_{1:t+1}$ the person was in, the current one s_{t+1} last; and the *actions* $a_{1:t}$ they took. Eq. 1 predicts a_{t+1}, the action in the current situation, from all four. *Memory* is not a fifth part but a summary of the four that a system keeps because it cannot hold the whole stream.

How each collapses into it. Each field's equation is Eq. 1 with the stream cut down to the part that field keeps:

$$\text{Rec.}\ \ p(i_{t+1} \mid i_{1:t}) = p(a_{t+1} \mid a_{1:t}) \tag{8}$$

$$\text{HCI}\ \ p(a_{t+1} \mid m_u,\, c_t) = p(a_{t+1} \mid g(h_u),\ s_{t+1}) \tag{9}$$

$$\text{Sim.}\ \ p(y \mid s,\, \pi_u) = p(a_{t+1} \mid \pi_u,\ E,\ s_{t+1}) \tag{10}$$

Left of each equals sign is the field's formula as §2 wrote it; right of it is Eq. 1 with the stream of Eq. 7 cut down to the parts shown. The equals sign means the same model with the stream restricted: the sides range over different events, items, screen actions, responses, and the identity is one of form. Recommendation keeps the actions: the items are the actions, and the environment, the catalog they are drawn from, is fixed; persona and situation are dropped, or partly tokenized: OneRec feeds user identity, age, and gender through a static pathway (Kuaishou OneRec Team, 2025), PLUM's prompt carries channel, watch time, and time since the last watch (He et al., 2025), and HSTU interleaves contextual features with actions (Zhai et al., 2024). HCI keeps the current situation, the screen $c_t = s_{t+1}$, and summarizes everything else into the memory $m_u = g(h_u)$. Simulation keeps the current situation s and the environment, and has the persona π_u supplied by the experimenter, a profile or an interview, rather than computed from a record; the response y is the action. A language model is Eq. 8 with words as actions.

What this shows. Each of Eqs. 8–10 is Eq. 1 with parts of the stream removed, summarized, or supplied, and that is the only way the fields differ. That difference is memory and context engineering, which §1.2 sets aside; with an unbounded stream and memory solved, the four are the same equation. When trained, each field fits Eq. 1 on the next event, by maximum likelihood in the recommenders, the same fit a language model uses, or by a reward scored on the predicted event in LongNAP and HumanLM; the prompted routes evaluate it without training. The loss each paper actually writes, sampled, masked, or reward-weighted, is in Table 4. The collapse is an identity of form, not a result, and the paper claims nothing more from it. What it does not show is that the fields learn the same representation of the person. Whether the cut-down stream costs anything is the empirical question the rest of the paper turns on.

3.2 The delta over a base language model

Table 1 collects, one measurement per field, what grounding the model in a person's own data adds over the base model's prior. Each row has its own metric, so the table ranks nothing across rows. The deltas are real in every field, smallest where the target is a stated response the base model can already approximate, and largest where the target is a revealed action on a real log. None of these measurements says whether the gain generalizes beyond the task it was measured on. No evaluation of the three fields' models on one task set exists; §4.1 says what it would need.

3.3 Scaling laws

Language loss falls as a power law in parameters, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Where behavioral papers publish fitted exponents, the numbers are close to the language ones. Compute-optimal allocation on 115 billion medical events gives $N \propto C^{0.520}$ and $D \propto C^{0.512}$, against 0.49 and 0.51 for text (Epic Cosmos and Microsoft Research, 2025; Hoffmann et al., 2022); a sequential recommender's loss falls faster in model size than a language model's, exponent 0.121 against 0.076 (Zhang et al., 2024b; Kaplan et al., 2020); a self-

Field	System	Base LLM	Grounded / trained	Metric
HCI	LongNAP	prompting	+39% (one user); +13% (new users)	judged similarity
Sim.	Interview agents	demographics, 0.74	interview in context, 0.83	normalized accuracy
Sim.	HumanLM	prompted, imitative	+16.3%	response alignment
Rec.	PLUM	LLM init, 0.23	behavioral pretraining, 0.28	recall@10
Rec.	AlignUSER	GPT-4.1, 21.5%	trained 8B, 52.9%	next-action accuracy

Table 1: Gain over a base language model, one row per measurement and one metric per row. Sources: Shaikh et al. (2026); Park et al. (2024); Wu et al. (2026); He et al. (2025); Bougie et al. (2026).

published calibration on event sequences tilts the allocation toward parameters, 0.617 and 0.383, with the data-to-parameter ratio falling from 344 to 36 as compute grows (Brüel Gabrielsson, 2026b). Several more report the power law without publishing an exponent: HSTU across three orders of compute, OneRec from 0.015B to 2.6B parameters, CLUE on 50 billion behavior tokens, Wukong past 100 GFLOP per example, and SimBench's fidelity log-linear in model size (Zhai et al., 2024; Kuaishou OneRec Team, 2025; Shin et al., 2023; Zhang et al., 2024a; Hu et al., 2025). Two qualifications. Scaling is task-dependent: Netflix fits an offset power law per task and finds some tasks near their fitted ceiling within the observed range while others keep improving (Xu et al., 2026). And every fitted law carries an irreducible term (Hoffmann et al., 2022), which for a person is the person's own inconsistency, the ceiling Park et al. (2024) already divide by. Scale is not what is missing.

3.4 Data composition

For language the corpus existed before the models did (§1.3). For behavior it does not, and Table 2 shows the shape of what exists: the corpora with consent see tens or thousands of people, and the corpora with millions see one platform or one record type. A pretraining corpus for language is organized by document; a person's traces are scattered across it, unlinked, and never presented as a sequence, which is why an LLM meets D3 only in context. Every behavioral pipeline begins by undoing this, assembling person-major sequences (Lei et al., 2026). The unit of the language corpus is the document; the unit of the behavioral corpus must be the person.

4 Direction from here

The goal of §1.2 is a computer that understands people. §3 showed three fields fitting one model to three slices of a person's stream, which is the case for unifying them: the slices belong to the same equation, so the data that produces them belongs in one corpus and one model. Architecture, objective, and scaling are settled to the standard of the other fields. The corpus and the evaluation are not, and the directions below are the steps from the three slices back to the whole.

4.1 Data

No corpus in Table 2 is longitudinal, cross-context, population-scale, and consented at once. Four things follow.

- *Breadth.* The same person observed across contexts. Within a company the join is free and has been published (Yuan et al., 2020, 2022); across companies no join exists. The alternative is the consented panel: people who carry their own streams across platforms, as GUM's participants did on their own devices (Shaikh et al., 2025). A panel of thousands is enough to measure cross-context transfer, not to pretrain on.
- *Depth.* Long, rich per-person sequences. The evidence is consistent across fields: user-representation quality grows smoothly with input sequence length (Shin et al., 2023), information gain is non-decreasing in history length (Jiang et al., 2026), the industrial frontier is lifelong sequences at lengths in the thousands (Pi et al., 2019; Zhai et al., 2024), and shuffling a month of history costs a next-action predictor accuracy (Shaikh et al., 2026). The trade between depth and breadth at a fixed budget has not been measured.
- *Consent.* A behavior stream identifies its author (Narayanan and Shmatikov, 2008), so de-identification is not a privacy mechanism for this data; consent at collection and the person's ability to read and revise the model (Shaikh et al., 2025) are.
- *Evaluation.* $G(z)$ is undefined until the task set is fixed (§1.2). Each field has a candidate benchmark (Zhou et al., 2025; Zhang et al., 2023; Hu et al., 2025; Huang et al., 2026; Toubia et al., 2025); none spans fields, and only the interview-agent protocol grades against the person's own consistency (Park et al., 2024). A standing benchmark needs both.

4.2 A common language for the stream

Eq. 7 names four parts. A model can train across the holders of a person's data only if all four arrive in one token language, and that is the mission this paper's directions reduce to: every holder tokenizes its slice of the stream, the tokens connect, and one model serves the fields' tasks. Today each field has solved one part its own way. Recommendation tokenized the action and is starting on the situation (§2.3); HCI captions the screen into language; simulation writes the persona and the situation as text. Language is the common denominator they are drifting toward, and OpenOneRec's interleaving of item codes with text is the first model that reads two of the parts in one context (Zhou et al., 2025).

The precedents say a shared vocabulary is what lets a model cross walls. Medicine got cross-country transfer once every event was an ICD code (Shmatko et al., 2025); pharma federated once every molecule was one fingerprint (Heyndrickx et al.,

Corpus	People	Contexts covered	Span per person	Access
GUM (Shaikh et al., 2025)	18	everything on one device	days	research consent
LongNAP (Shaikh et al., 2026)	20	everything on one screen	one month	research consent
Interview agents (Park et al., 2024)	1,052	a whole life, elicited	two hours, retest at two weeks	research consent
Twin-2K-500 (Toubia et al., 2025)	2,058	500 questions, elicited	four waves	research consent
SocSci210 (Kolluri et al., 2025)	400K	210 experiments, elicited	one session	research consent
HumanLLM traces (Lei et al., 2026)	282K accounts	public posts, four platforms	account history	public, unconsented
Tenrec (Yuan et al., 2022)	5M	four surfaces, one company	log window	open release
Yambda (Yandex, 2025)	1M	one music platform	11 months	open release
MicroLens (Ni et al., 2025)	34M	one video platform	log window	open release
Kuaishou (Kuaishou OneRec Team, 2025)	platform scale	one video platform	lifelong log	private
life2vec (Savcisens et al., 2024)	6M	health, work, income, residence, education	ten years	registry, restricted
CoMET (Epic Cosmos and Microsoft Research, 2025)	118M	health records, 310 systems	patient history	private
Delphi-2M (Shmatko et al., 2025)	0.4M + 1.9M	disease history	decades	biobank, registry

Table 2: The corpora the surveyed models train on. No row has many people, many contexts of a person's life, a long span, and research consent at once.

2024). Behavior has no such standard for any of its four parts. The nearest existing piece is the semantic ID: a tokenizer built on a public content encoder gives every silo a common code for the action, each silo tokenizes locally, and what leaves is a sequence of codes rather than a log (Rajput et al., 2023). The persona, environment, and situation have no equivalent yet.

Once a language exists, two routes to training across holders are visible. Federated training: MELLODDY had ten competing pharmaceutical companies pretrain jointly without sharing data, and each improved (Heyndrickx et al., 2024); federated pretraining has been demonstrated for language models (Sani et al., 2024) and not yet for behavioral ones. Translation between representations: the user embeddings each silo already holds could be mapped into one space rather than retrained from pooled data, on the evidence that representations of scaled models converge (Huh et al., 2024). Both need the language first.

4.3 Learning from recommenders

§2.3 ended on a one-way flow. PLUM states the forward problem: language models are not pretrained on user behavior, so adapting them to recommendation means teaching them a domain (He et al., 2025). The framing takes for granted that behavior is a domain to be taught, never a modality to learn from. Nothing in the recipe forbids the reverse: behavioral tokens can share a context with text (Zhou et al., 2025), a person can enter a language model as a learned embedding rather than a prompt (Ning et al., 2024), and a recommender's grounding is revealed rather than stated preference, the signal §1.4 says text lacks. Why has no language model been shown to gain knowledge of people from behavior? Either nobody measured it, the behavioral data is too narrow, the knowledge forms but has no path into the language circuits, or nothing person-shaped forms at all. The first is the cheapest to eliminate: OpenOneRec ships checkpoints with base models of the same sizes, and the simulation benchmarks are the person-task suites. If the transfer is zero, behavioral models remain a parallel line; if it is positive, behavioral data is the pretraining substrate that follows text.

5 Discussion

Domain specialists. This is the paper's point, so it is worth stating in full. The field that has the data, population-scale, longitudinal, revealed behavior, produces specialists: the leading open recommender describes its own field as constrained by isolated data and operating as "domain specialists" (Zhou et al., 2025). The fields that show general behavior, an agent answering a survey it never saw, an assistant predicting an action on a new app, get it from text: they post-train a language model, and the generality is the language model's (§3). So the two halves of a foundation model of human behavior both exist, and they are held apart. Recommendation has behavioral pretraining and lacks breadth; HCI and simulation have breadth and lack behavioral pretraining. Neither can reach the other with what it has. A recommender cannot become general by scaling on one platform's log, because generality across a person's contexts requires observing those contexts (§3.4), and a language model cannot become behavioral by prompting, because the per-person stream is not in text (§1.4, Table 1). A foundation model of human behavior is what results when behavioral pretraining at recommendation's scale runs on a stream that spans the contexts the other two fields work in. The specialist reading, that behavior is bound to its context and a general model of a person is a category error, predicts the same present as this paper does; the two predict different futures, and only a cross-context corpus in a common language (§4) separates them. Put simply: specificity buys prediction, a foundation model buys generalization, and the way to have both is one token language for the stream and one model trained on all of it.

The goal. None of this is for its own sake. The aim set in §1.2 is a computer that understands a person, meaning one that can predict or simulate what they will do in any situation, and a universal representation matters only as the means to that: the

common language and the corpus are how the stream that test requires gets assembled.

Text sufficiency. The competing null is that a frontier language model with the right context is all that is required. It holds for population-level and stated-response tasks and fails for per-person revealed behavior (Table 1), and the boundary between the two is itself the measurement the field should make.

Limitations of this survey. Three fields were chosen, and health, finance, advertising, and mobility enter only where they bear on scaling and transfer. Two of the three fields' leading systems come from one research group, so some of the agreement reported is shared practice rather than independent arrival. Most industrial results are self-reported by the platforms that built the systems and have not been replicated outside them, and one scaling study is a company blog post. The reduction of §3.1 is an identity of form, and the deltas of Table 1 are measured on incommensurate metrics, so the cross-field comparison is qualitative. There is also a standing argument that generalization in recommendation is limited by nature: that behavior on one platform, in one catalog, under one interface, is too bound to its context to carry over, and the working optimum is a specialist retrained often. The evidence against it is that a model pretrained on one platform's logs transfers to ten catalogs it never saw, at $+26.8\%$ average Recall@10 (Zhou et al., 2025), which is generalization across contexts of the item if not yet of the person; we read the objection as a claim about today's corpora rather than about behavior, and §4.3 is where it gets tested in the other direction. Much of the cited work is a 2025 or 2026 preprint; every claim was checked against its source, and the field moves faster than a survey does.

Risks. The representation that serves a person is the representation that profiles them: private traits are predictable from records of Likes (Kosinski et al., 2013), targeting built on such predictions moves behavior (Matz et al., 2017), and language models infer personal attributes from text (Staab et al., 2024). Generality makes profiling a side effect rather than a design choice, so the governance question is who runs the model and under what consent.

6 Conclusion

We began with a goal that psychology, cognitive science, sociology, anthropology, and economics have each pursued and none has reached: a computer that understands people, meaning one that can predict or simulate what a person will do in any situation. Deep learning reached goals of that kind in other domains once three things were in place: a generic objective, a model that scales under it, and a corpus large enough to scale on. For language the three arrived together, and understanding fell out.

This paper's finding is that for people, two of the three are already in place. The objective is next-action prediction over a person's stream, Eq. 1, and three fields have been fitting it without saying so, each with the stream cut down to the part it can see (§3.1). Scaling holds wherever the data is dense, with exponents close to the language values (§3.3). The third thing is missing, and it is missing in a specific way. The population-scale data sits inside single platforms and produces domain specialists; the cross-context data sits with tens or thousands of consenting people and is too small to pretrain on; and no common token language joins them (§3.4, §4).

That is the vector. The claim is not that a model of people is assured. It is that, for the first time, the problem has the shape of the ones deep learning has solved, and that what stands between the present and a foundation model of human behavior is not an objective or an architecture but a corpus. Data is the most important ingredient, and for behavior it is the one that has to be built rather than found.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models, 2024. arXiv:2402.16827.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12): 29–38, 1992.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5185–5198, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. arXiv:1206.5538.

Marcel Binz, Eric Schulz, et al. A foundation model to predict and capture human cognition. *Nature*, 644:1002–1009, 2025. arXiv:2410.20268.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint*, 2021. arXiv:2108.07258.

Nicolas Bougie, Gian Maria Marconi, Xiaotong Ye, and Narimasa Watanabe. AlignUSER: Human-aligned LLM agents via world models for recommender system evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 16433–16451, 2026.

Rickard Brüel Gabrielsson. Large behavioral models: A foundation model paradigm for human actions, 2026a. URL https://research.unboxai.com/large-behavioral-models.pdf.

Rickard Brüel Gabrielsson. Scaling laws for behavioral foundation models over user event sequences, 2026b. URL https://research.unboxai.com/scaling-laws-for-behavioral-foundation-models.html.

Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, et al. OneSearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search, 2025. arXiv:2509.03236.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 2016. arXiv:1606.07792.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory, 2025. arXiv:2504.19413.

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, 2016.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience, 2026. arXiv:2605.04326.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models, 2024. arXiv:2306.07951.

Epic Cosmos and Microsoft Research. Generative medical event models improve with scale, 2025. arXiv:2508.12104.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys)*, 2022. arXiv:2203.13366.

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3:833–838, 2023.

Ruining He, Lukasz Heldt, Lichan Hong, Raghunandan Keshavan, Shifan Mao, Nikhil Mehta, Zhengyang Su, Alicia Tsai, Yueqi Wang, Shao-Chuan Wang, Xinyang Yi, et al. PLUM: Adapting pre-trained language models for industrial-scale generative recommendations, 2025. arXiv:2510.07784.

Zhicheng He, Weiwen Liu, Wei Guo, Jiarui Qin, Yingxue Zhang, Yaochen Hu, and Ruiming Tang. A survey on user behavior modeling in recommender systems. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23), Survey Track*, pages 6656–6664, 2023.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models, 2024.

Wouter Heyndrickx, Lewis Mervin, Tobias Morawietz, et al. MELLODDY: Cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. *Journal of Chemical Information and Modeling*, 64(7):2331–2344, 2024.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. arXiv:1511.06939.

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. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2203.15556.

Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts LLM performance. Technical report, Chroma, 2025.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus?, 2023. arXiv:2301.07543.

Yupeng Hou, Haven Kim, Clark Mingxuan Ju, Eduardo Escoto, Neil Shah, and Julian McAuley. Expressiveness limits of autoregressive semantic ID generation in generative recommendation, 2026. arXiv:2605.06331.

Tiancheng Hu et al. SimBench: Benchmarking the ability of large language models to simulate human behaviors, 2025. arXiv:2510.17516.

Jin Huang, Yutong Xie, Wanli Song, Xingjian Zhang, Walter Yuan, Matthew O. Jackson, and Qiaozhu Mei. BehaviorBench: Benchmarking foundation models for behavioral science tasks, 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2405.07987.

Shali Jiang, Hua Zheng, Boyang Liu, Laming Chen, Kenny Lov, Chuanqi Xu, Lisang Ding, Qinghai Zhou, Can Cui, Xiaolong Liu, et al. LoopFM: Learning frOm HistOrical RePresentations of foundation model for recommendation. *arXiv preprint arXiv:2605.29280*, 2026. URL https://arxiv.org/abs/2605.29280.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining (ICDM)*, 2018. arXiv:1808.09781.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. arXiv:2001.08361.

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*, 2025. arXiv:2509.05830.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

Kuaishou OneRec Team. OneRec technical report, 2025. arXiv:2506.13695.

Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation. *Artificial Intelligence Review*, 2026.

Y. Lei, Y. Wang, Jianxun Lian, X. Hu, Defu Lian, and Xing Xie. HumanLLM: Towards personalized understanding and simulation of human nature. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. arXiv:2601.15793.

Ying Li, Shradha Sehgal, and Arjun Rao. GenRec: An LLM-backed recommendation ranker at Netflix, 2026. arXiv:2608.10257.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

Sandra C. Matz, Michal Kosinski, Gideon Nave, and David J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48):12714–12719, 2017.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, 2008.

Netflix Technology Blog. Foundation model for personalized recommendation. https://netflixtechblog.medium.com/, 2025.

Allen Newell. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA, 1990.

Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. A content-driven micro-video recommendation dataset at scale. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2025. arXiv:2309.15379.

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. USER-LLM: Efficient LLM contextualization with user embeddings, 2024. arXiv:2402.13598.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2203.02155.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023. arXiv:2310.08560.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *International Conference on Learning Representations*, 2024. arXiv:2309.05196.

Nikil Pancha, Andrew Zhai, Jure Leskovec, and Charles Rosenberg. PinnerFormer: Sequence modeling for user representation at Pinterest. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. arXiv:2205.04507.

Joon Sung Park, Lindsay Popowski, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2022. URL https://arxiv.org/abs/2208.04024.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. arXiv:2304.03442.

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

Qi Pi, Weijie Bian, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Practice on long sequential user behavior modeling for click-through rate prediction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019. arXiv:1905.09248.

Matteo Pirotta, Andrea Tirinzoni, Ahmed Touati, Alessandro Lazaric, and Yann Ollivier. Fast imitation via behavior foundation models. In *International Conference on Learning Representations (ICLR)*, 2024.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, et al. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.05065.

RecGPT Team. RecGPT-V3 technical report, 2026. arXiv:2607.15591.

Paul Resnick and Hal R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

Elaine Rich. User modeling via stereotypes. *Cognitive Science*, 3(4):329–354, 1979.

Lorenzo Sani, Alex Iacob, Zeyu Cao, Bill Marino, Yan Gao, Tomas Paulik, Wanru Zhao, William F. Shen, Preslav Aleksandrov, Xinchi Qiu, and Nicholas D. Lane. The future of large language model pre-training is federated, 2024. arXiv:2405.10853.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 2023. arXiv:2303.17548.

Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. Using sequences of life-events to predict human lives. *Nature Computational Science*, 4:43–56, 2024.

Sarah Schröder et al. Large language models do not simulate human psychology, 2025. arXiv:2508.06950.

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980.

Omar Shaikh, Shardul Sapkota, et al. Creating general user models from computer use. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2025. arXiv:2505.10831.

Omar Shaikh et al. Learning next action predictors from human-computer interaction, 2026. arXiv:2603.05923.

Kyuyong Shin, Hanock Kwak, Su Young Kim, Max Nihlen Ramstrom, Jisu Jeong, Jung-Woo Ha, and Kyung-Min Kim. Scaling law for recommendation models: Towards general-purpose user representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4596–4604, 2023. URL https://arxiv.org/abs/2111.11294.

Kyuyong Shin et al. One4all user representation for recommender systems in e-commerce, 2021. arXiv:2106.00573.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, et al. Learning the natural history of human disease with generative transformers. *Nature*, 2025.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.07298.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019. arXiv:1904.06690.

Shixiang Tang, Yizhou Wang, Lu Chen, Yuan Wang, Sida Peng, Dan Xu, and Wanli Ouyang. Human-centric foundation models: Perception, generation and agentic modeling. *arXiv preprint*, 2025. arXiv:2502.08556.

Olivier Toubia et al. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions, 2025. arXiv:2505.17479.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks, 2023. arXiv:2302.08399.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of LLM scaling

based on human-generated data. *arXiv preprint arXiv:2211.04325*, 2024.

Shirley Wu et al. HumanLM: Simulating users with state alignment beats response imitation, 2026. arXiv:2603.03303.

Qiuling Xu, Ko-Jen Hsiao, and Moumita Bhattacharya. Towards generalizable and efficient large-scale generative recommenders, 2026. arXiv:2605.23312.

Yandex. Yambda-5B: A large-scale multi-modal dataset for ranking and retrieval, 2025. arXiv:2505.22238.

Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. arXiv:2001.04253.

Guanghu Yuan et al. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. arXiv:2210.10629.

Mingqi Yuan, Tao Yu, Wenqi Ge, Xiuyong Yao, Huijiang Wang, Jiayu Chen, Bo Li, Wei Zhang, Wenjun Zeng, Hua Chen, and Xin Jin. A survey of behavior foundation model: Next-generation whole-body control system of humanoid robots. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. arXiv:2506.20487.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv:2402.17152.

Buyun Zhang, Liang Luo, Yuxin Chen, Jade Nie, Xi Liu, Shen Li, Yanli Zhao, Yuchen Hao, Yantao Yao, Ellie Dingqiao Wen, et al. Wukong: Towards a scaling law for large-scale recommendation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

Gaowei Zhang, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. Scaling law of large sequential recommendation models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024b. arXiv:2311.11351.

Jiaqi Zhang et al. NineRec: A benchmark dataset suite for evaluating transferable recommendation, 2023. arXiv:2309.07705.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1059–1068, 2018.

Guorui Zhou et al. OpenOneRec technical report: An open foundation model and benchmark to accelerate generative recommendation, 2025. arXiv:2512.24762.

Xuhui Zhou et al. OdysSim: Building foundation models for human behavior simulation, 2026. arXiv:2606.14199.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008.

Caleb Ziems, Dora Zhao, Rose E. Wang, Matthew Jörke, James others (58 authors, incl. Landay, and Diyi) Yang. Reflections and new directions for human-centered large language models, 2026. arXiv:2605.06901.

Yanyan Zou, Junbo Qi, Lunsong Huang, et al. GenRec: A preference-oriented generative framework for large-scale recommendation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2026. arXiv:2604.14878.

A Other names for the same model

The same ambition travels under many names, coined in communities that have not reconciled their vocabularies. Table 3 collects them. Rows above the rule are near-synonyms for what this paper means by a foundation model of human behavior, a model of a person learned from behavioral traces. Rows below are false friends: they share the words but model something else (an agent's control policies, a digital human's appearance and motion, the brain's response to stimuli, an environment's dynamics).

Term	Community	What it models
Foundation model of human behavior	this paper	a person, from behavioral traces
Large Behavioral Model (LBM)	industry	human action and event sequences
Foundation model of human cognition	cognitive science	trial-level cognition (Centaur)
Generative agent	HCI / simulation	believable individual behavior (prompted)
General user model (GUM)	HCI / assistants	a user's context, goals, and state
User behavior modeling (UBM)	recsys / industrial	user interest from interaction sequences
Digital twin (of a person)	simulation / industry	a specific person's responses
Silicon sampling	comp. social science	population survey responses
False friends: same words, different thing		
Behavior Foundation Model (BFM)	RL / robotics	an agent's control policies
Human-Centric Foundation Model	computer vision	digital-human appearance and motion
Foundation model of the brain	neuroscience	neural responses to stimuli (TRIBE)
World model	model-based RL	environment dynamics, not the person

Table 3: Other names for the same model. Representative works: Centaur (Binz et al., 2025); LBM (Brüel Gabrielsson, 2026a); generative agents (Park et al., 2023); GUM (Shaikh et al., 2025); UBM (He et al., 2023); digital twins (Toubia et al., 2025); silicon sampling (Argyle et al., 2023); BFM (Pirotta et al., 2024; Yuan et al., 2025); HCFM (Tang et al., 2025); TRIBE (d'Ascoli et al., 2026).

A.1 Stated definitions

For simplicity the main text adopts one definition; the source definitions it was weighed against are collected here.

Foundation model (Bommasani et al., 2021): "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks."

Foundation model of human cognition (Binz et al., 2025): "a computational model that can predict and simulate human behavior in any experiment expressible in natural language."

General user model (Shaikh et al., 2025): a model that infers a user's context, goals, and state from observations of their computer use, maintained as natural-language propositions with confidence levels.

Large Behavioral Model (Brüel Gabrielsson, 2026a): "foundation models where the co-occurrence within chronological sequences of human behaviors and actions defines the meaning of entities and features acted upon."

Behavior Foundation Model (the RL false friend) (Pirotta et al., 2024): a pre-trained model of an agent's own control policies, adaptable zero-shot to new tasks; the behavior is the robot's, not a person's.

Industrial labs also describe single-domain behavioral models as foundation models (Netflix Technology Blog, 2025); under the definition of §1.4 these satisfy D1–D3 within one platform's traces but meet D4 only across that platform's surfaces.

B The formulas the fields wrote

The predictor of Eq. 1 was not invented here; the fields wrote it repeatedly, in their own notation. Table 4 collects the objectives as the papers state them (long conditioning variables abbreviated: Ctx, Hist). The recommendation entries fall into four families, sampled discriminative, raw-ID sequential transduction, semantic-ID generation, and reward-weighted or preference-aligned; the families differ in the event (raw ID, codeword, interleaved token), the factorization (autoregressive or masked), and the weighting (uniform, reward, or policy gradient). Every trainable entry fits Eq. 1 by maximum likelihood; every entry is Eq. 1 with a particular stand-in for the stream.

System	Field	Objective, as written	Person enters as
Rasch / IRT (Lord and Novick, 1968)	psychometrics	$P(\text{correct} \mid \theta_u, b_i) = \sigma(\theta_u - b_i)$	scalar ability θ_u
MaxEnt IRL / BToM (Ziebart et al., 2008; Baker et al., 2009)	RL / cog. sci.	$P(a \mid s, R_u)$	reward, beliefs R_u
Matrix factorization (Koren et al., 2009)	recommendation	$P(r \mid u \cdot i)$	learned vector u
Language model (Bommasani et al., 2021)	NLP	$-\sum_t \log p_\theta(w_{t+1} \mid w_{\le t})$	latent in weights
SASRec (Kang and McAuley, 2018)	recommendation	$-\sum_t \left[\log \sigma(r_{o_t,t}) + \sum_{j \notin S_u} \log(1 - \sigma(r_{j,t}))\right]$	hidden state
BERT4Rec (Sun et al., 2019)	recommendation	Cloze: predict masked items from both directions	hidden state
HSTU (Zhai et al., 2024)	recommendation	$p(a_{i+1} \mid \Phi_0, a_0, \Phi_1, a_1, \ldots, \Phi_{i+1})$	hidden state
TIGER (Rajput et al., 2023)	recommendation	seq2seq NLL over the next item's semantic-ID codewords	hidden state
OneRec (Kuaishou OneRec Team, 2025)	recommendation	$\mathcal{L}_{\text{NTP}} = -\sum_{j=0}^{L_t-1} \log P\big(s_m^{j+1} \mid s^{[\text{BOS}]}, s_m^1, \ldots, s_m^j\big)$, then on-policy RL (preference, format, industrial rewards)	hidden state
OpenOneRec (Zhou et al., 2025)	recommendation	next-token NLL over interleaved text and itemic tokens	hidden state
PLUM (He et al., 2025)	recommendation	$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{L} r(u, v_{\text{click}}) \cdot \log P(\text{sid}_t \mid \text{Ctx}_u, \text{Hist}_u, \text{sid}_{<t})$	context + history tokens
GenRec (Zou et al., 2026)	recommendation	page-wise next-token NLL, then GRPO over hybrid rewards	hidden state
life2vec (Savcisens et al., 2024)	life events	masked event modeling (30%) + sequence-order prediction	person embedding
Silicon sampling (Argyle et al., 2023)	simulation	$p_{\text{LLM}}(y \mid s, \pi_u)$, π_u a demographic backstory	supplied persona
Interview agents (Park et al., 2024)	simulation	evaluation: $\text{acc}_{\text{agent}}/\text{acc}_{\text{self-replication}}$ (normalized accuracy)	interview transcript
HumanLM (Wu et al., 2026)	simulation	$p_\theta(y \mid p, x)$ over $\{(p^{(i)}, x^{(i)}, y^{(i)})\}$; GRPO on judged latent-state alignment	supplied profile p
Centaur (Binz et al., 2025)	cognition	cross-entropy, masked to human-response tokens (QLoRA)	weights + trial context
GUM (Shaikh et al., 2025)	HCI	confidence-weighted propositions, no trained loss	text propositions m_u
LongNAP (Shaikh et al., 2026)	HCI	GRPO maximizing LLM-judged similarity of predicted future actions; retrieval decayed by $e^{-\lambda \text{ age}}$	per-user memory + weights
USER-LLM (Ning et al., 2024)	recommendation / LLM	cross-attention on a pretrained user-encoder embedding	latent embedding
RecGPT (RecGPT Team, 2026)	recommendation / LLM	$p_\theta(z, y \mid \text{persona}, \text{history})$ with curated memory	written persona + memory

Table 4: The objectives the fields actually wrote, transcribed from the papers. Every trainable row fits Eq. 1 by maximum likelihood; the rows differ in the event, the factorization, the weighting, and what stands in for z_u.