BehaviorGPT at Work: A Foundation Model for Workforce Actions & Dynamics

Rickard Brüel Gabrielsson[*] **Vasudev Gupta, et al.**
Unbox AI

Abstract

We applied language modeling techniques to detailed workforce behavioral data, creating BehaviorGPT-v2: a foundation model treating employee behaviors as a language to predict future actions and outcomes. We expand BehaviorGPT into a foundation model for understanding and predicting workforce behaviors and dynamics. We apply language modeling principles to employee and employer interactions by representing behaviors as sequences of tokens, effectively capturing the implicit "language" of workforce dynamics. Using this approach, our model achieves 91% accuracy (F1 score: 0.93) in predicting employee attrition (whether an employee would quit within the next month and why). This represents a substantial improvement over traditional data analyses and survey-based methods, which previously provided minimal predictive value, clearly demonstrating that actions speak louder than words. Our results suggest employee-employer interactions contain semantic structure akin to language, and thus can be modeled effectively. Specifically, we train a Transformer-based model variant on a dataset of approximately 43 million behavioral events involving 80,000 employees over a four-year period. Although dense embeddings were generated for visualization and analytical purposes, attrition predictions are performed directly through end-to-end modeling, analogous to predicting words in language models. Crucially, BehaviorGPT is not specifically optimized for a particular downstream task, but rather trained with the guiding principle: "predicting what you will do, means understanding who you are." Building upon our earlier consumption and transaction-focused work, we now model employers as "merchants" and employees as their "customers," interpreting their interactions through this behavioral language.

Highlights:

- We demonstrate that workforce actions constitute a predictable implicit language, significantly outperforming traditional survey-based and analytic methods.
- We achieve 91% accuracy in predicting whether employees would quit within the following month.

1 Introduction

Foundation models are large neural networks trained on massive datasets to learn general-purpose representations. They have shown remarkable success in domains such as language, vision, and genomics. Their effectiveness lies in capturing complex patterns from diverse data before fine-tuning or prompting for specific tasks. Yet the majority of human-generated data isn't textual or visual, it is behavioral. Every day, billions of individuals generate vast sequences of discrete actions, including

[*] Corresponding author: rickard@unboxai.com

Original post (published June 26, 2025): `https://research.unboxai.com/behaviorgpt-foundation-model-workforce.html` transactions, online interactions, and workplace tasks. Modeling behavioral data represents the next frontier in foundational AI capabilities.

In language modeling, words (tokens) appear sequentially, and models predict subsequent tokens by distilling rich linguistic and contextual relationships. Similarly, behavioral data consists of chronological event streams, such as user sessions on retail platforms, payment transactions, or employee shift records. By applying a next-event prediction paradigm to large-scale behavioral datasets, we can generate powerful representations of both the individuals (agents) and their interactions (objects or services).

For many companies, employee salaries and related expenses dominate their costs, estimated at about 70% (Paycor labor-cost analysis). Employees also drive most organizational value creation. Thus, improved understanding of employee behaviors and interactions offers significant benefits for both employees and employers. Attrition, the loss of employees, is particularly costly. The company we partnered with received applications from approximately 500,000 candidates annually but faced an annual turnover rate of about 70%: out of 30,000 employees, roughly 20,000 departed each year. Many left before they could generate substantial value, with an average tenure of just 1.5 years. While this attrition problem initially motivated our research, our broader goal was to understand employees more fundamentally through their actions and behaviors.

A deep understanding of employee behavior can benefit organizations by:

- Improving scheduling and operational efficiency (who should work, and when).
- Identifying and recruiting individuals who will remain engaged and provide value.
- Developing targeted motivational and retention strategies.
- Reducing costs associated with high employee turnover.
- Recognizing behavioral signs indicating belonging, burnout, or misalignment, enabling timely support and interventions.

Comparing our bottom-up foundation modeling approach with previous top-down surveys and traditional analytic methods illustrates vividly that "actions speak louder than words." Modeling employees' actual behaviors dramatically outperformed conventional approaches, indicating a clear shift toward behavioral foundation models as the future of actionable business insights.

Finally, we acknowledge the ethical complexities inherent in tracking and analyzing employee behaviors. Employee-employer relationships are sensitive, making responsible data handling and transparent use crucial. This research was conducted strictly for exploratory purposes, and its real-world implementation warrants extensive ethical consideration. Highlighting the sensitivity, one team member chose to leave the project due to ethical discomfort of studying employee behavior and the sensitive relationships between employees and employers, underscoring the importance of careful consideration and dialogue around these powerful modeling capabilities.

2 Where is the “language” of workforce behavior?

Not all data is created equal. To uncover a behavioral “language” in workforce interactions, where patterns can be modeled and predicted, we need more than just large volumes of data. We need meaningful and structured signals. This section outlines the key criteria for identifying such a language within workforce data.

2.1 Meaningful and predictive signals

A foundation model can only learn from data if that data contains inherent structure. While it's possible to feed the model large volumes of data and allow it to identify patterns, meaningful signals must be present for learning to occur. Pure noise offers nothing to compress or predict. If the future is statistically independent of the past (i.e., $p(\text{future} \mid \text{history}) = p(\text{future})$), no model can uncover useful patterns. Thus, we focus on behavioral events that are consistent indicators of individual traits—actions that reveal preferences, motivations, and likely future behavior.

We find that survey responses and self-reported forms, while sometimes useful for diagnostics, are often poor predictors of real-world behavior due to their inconsistency and low signal-to-noise ratio. In contrast, logged behavioral events—that is, how people spend time, take breaks, change schedules—are more reliable and predictive.

2.2 Volume and variety of events

For a foundation model to generalize, it must observe enough high-quality examples. We need tens of millions of relevant historical sequences where future behavior can be predicted from past activity. Crucially, these sequences must contain consistent event types. If each user exhibits only unique types of events (e.g., one user only takes breaks, another only has sick days), the model struggles to learn generalizable patterns across employees.

It is more valuable to have co-occurring and overlapping event types (e.g., how breaks and sick days might jointly predict vacation time) than to have disjoint sets of behaviors per user. In short, we aim to capture joint probabilities across event types, not isolated patterns.

2.3 Depth over breadth

We benefit more from deeply understanding a few individuals than from shallow glimpses of many. Long sequences of meaningful events per employee enable the model to infer latent traits and behavioral tendencies, which are foundational for making accurate predictions about unseen users.

2.4 Analogies in human natural language

Human natural language exhibits all of these desirable properties. Because language was developed as a tool for communication, we know it is not random; it carries structured, meaningful information, and it is inherently learnable (as evidenced by our ability to acquire it). Language is composed from a finite set of symbols, such as words or characters, meaning that sequences of text share common elements that can act as building blocks for learning. Moreover, we have access to long, coherent sequences produced by individuals (such as books), which provide rich insight into what the author intends to communicate.

3 Our dataset

We curated a dataset of 43 million behavioral events from 80,000 employees—an average of 538 events per individual. These sequences include 15 event types, each tracked weekly and sorted chronologically. Key categories include:

- Scheduled working hours
- Actual working hours
- Productive hours
- Paid and unpaid breaks
- . . . plus 10 additional behavioral signals

In addition to event data, we incorporated contextual demographic features—non-event attributes that provide critical context (which we call "domains"), such as:

- Employee's manager
- Campaign or task assignment
- . . . plus 6 additional domains

All personally identifiable information was removed. Employee IDs were salted, and timestamps were jittered to preserve privacy while retaining sequence structure.

Importantly, data splits for training, validation, and testing were made by user, not by event, ensuring that model evaluation reflects its ability to generalize to unseen individuals rather than the easier task of merely generalizing to future actions of already seen individuals.

4 Architecture

We adopted an architecture closely aligned with our previous work in modeling grocery consumption behavior, with minimal modifications. The model follows a standard Transformer-based design, with a total of 5M parameters, adapted for sequential behavioral data.

4.1 Model components

- Event Embedding: Each behavioral event-type is one-hot embedded (i.e., each type has its own unique token). These are passed through an embedder to generate dense vector representations.
- Transformer Backbone: The embedded sequences are processed by an 8-layer Transformer, which models temporal dependencies and contextual relationships between events.
- Decoder Head: The output of the Transformer is decoded to generate predictions, such as the likelihood of future events (e.g., attrition) based on past behavior. A different multi-class classification head is used for fine-tuning on attrition tasks.

4.2 Training strategy

We employed a two-stage training procedure, mirroring our approach in consumption modeling. However, unlike product catalogs or search terms, which frequently introduce unseen events such as new products and search terms, the set of employee actions is more constrained and stable. As a result, generalization to entirely new event types is less critical in this context. Nevertheless, it is simple to use text descriptions (“break”, “overtime”, etc) as features such that the model retains some ability to adapt to novel or sparsely represented behaviors due to the shared embedding space.

4.3 Transferability

One of the most notable findings was the strong architectural transferability: the same model structure and hyperparameters that performed well in consumer modeling also yielded strong results in modeling workforce dynamics. This suggests that our approach of treating human behavior as a structured, sequential language is broadly applicable across domains, from consumption to labor and beyond. It is universally applicable wherever human actions are logged over time. It extends to domains such as education (e.g., student learning paths), healthcare (e.g., patient treatment decisions), finance (e.g., investment behavior), robotics (e.g., behavior of menial tasks), creative tools (e.g., sequences of edits in design), and much more.

Fundamentally, wherever individuals interact with structured systems over time, their actions reveal intent, skill, context, and trajectory, just as words reveal meaning in language. This universality opens the door to a unified approach to modeling real-world behavior across sectors, all grounded in the same modeling primitives: tokens, sequences, and temporal context.

5 Learn more by solving the hardest task

“Aim for the stars, hit the moon”

Traditional approaches to workforce prediction, such as attrition modeling, start with a narrowly defined, labeled dataset: sequences of employee behavior paired with labels indicating whether an employee quit. A model is then trained directly on this supervised task. While this can produce usable results, it limits the model’s learning capacity and generalization ability.

In contrast, foundation models and self-supervised learning follow a different philosophy: learn as much as possible about the structure of the data first, then specialize only if necessary. Rather than optimizing immediately for a single outcome like attrition, we pretrain the model on a broader, more challenging task: predicting the next event in a behavioral sequence, given all prior events. This is analogous to language modeling, where predicting the next word forces the model to understand syntax, semantics, and context, before being fine-tuned for downstream tasks like question answering or sentiment analysis.

By training on this harder, more general objective, the model first learns the “behavioral language” of employees, e.g. their routines, rhythms, and deviations. Only after this foundational knowledge is acquired do we consider fine-tuning or evaluating on specific tasks like attrition prediction. This approach ensures that the model develops a deep, transferable understanding of human behavior within organizations.

Crucially, this method is not just better for predicting who will leave, it also helps us understand *why* people leave, and more importantly, how to keep them engaged, satisfied, and productive.

6 Is workforce behavior compressible?

Before modeling workforce behavior, we must ask a foundational question: Is it predictable? In other words, is there enough signal in employee actions to compress, i.e., to represent the behavior in a lower-entropy form that still preserves structure and meaning?

In any real-world process, some randomness is inevitable due to unobserved variables. But the key question is whether the observable behavior contains sufficient structure to enable meaningful prediction. To measure this, we use cross-entropy loss which is a standard metric in sequence modeling. Conceptually, the exponential of the cross-entropy loss approximates the number of distinct action-patterns (or "tokens") the model needs to describe behavior accurately. Lower loss implies higher predictability and compressibility.

In Figure 1, we show both training and test loss curves during pretraining. The training loss decreases steadily, indicating that the model is learning to predict the next action based on prior behavior. Because our Transformer model is heavily parameterized (i.e., it has almost as many parameters as training examples) and we iterate over training data multiple times, one might worry about memorization.

However, the test loss also decreases and stabilizes at a low level, demonstrating that the model is not simply memorizing, but instead generalizing to new, unseen employee behavior sequences (i.e. data it was not allowed to see during training). This indicates that workforce behavior, at least as captured in our dataset, is genuinely predictable and compressible, which is a necessary condition for building a useful behavioral foundation model.

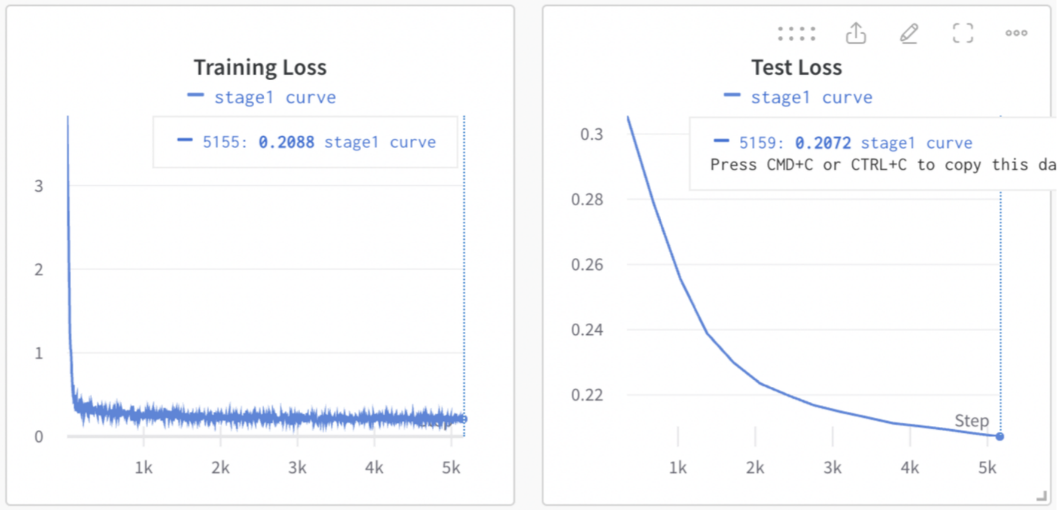

Figure 1: Cross-entropy loss during pretraining for next-event prediction in employee behavior sequences.

7 What does the model learn?

7.1 Exploring the embedding space

To understand what the model has learned during pretraining, we analyze the embedding space, i.e. the dense vector representations that summarize an employee's behavioral history.

Specifically, we run each employee's sequence of events through the pretrained Transformer and compute the mean of the hidden state embeddings. This results in a single 256-dimensional vector per employee that captures their behavioral signature. We then reduce this high-dimensional space to two dimensions using UMAP for visualization.

7.2 Country-level behavioral clustering

In Figure 2, we color each employee embedding by country of employment. Distinct clustering patterns emerge, indicating that country strongly correlates with behavioral patterns. This suggests that cultural, regulatory, or organizational norms tied to geography influence how employees behave in measurable ways.

However, the clusters are not strictly country-bound. The embedding space reveals more nuanced relationships, indicating that the model captures subtleties beyond just location, perhaps including work style, schedule adherence, or team dynamics.

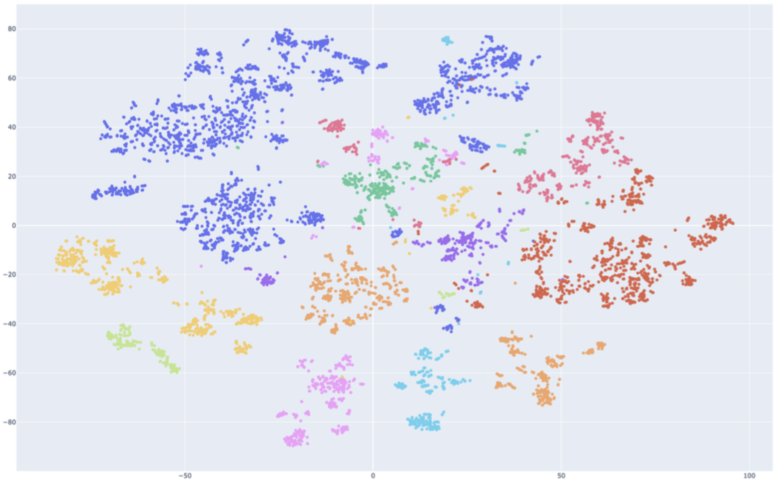

Figure 2: UMAP projection of employee embeddings, colored by country (e.g., United States, Germany, Philippines).

7.3 Reason for leaving

Next, we color embeddings based on reason for leaving (voluntary, involuntary, leave of absence, still employed, etc.). While there are some visible patterns and groupings, the correlation is less pronounced than with country, suggesting that reason-for-leaving is encoded in a more distributed or subtle manner within the embedding space.

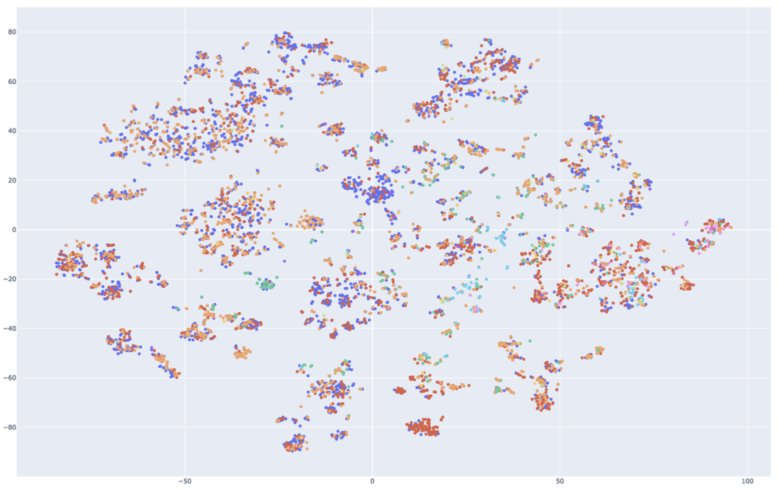

Figure 3: UMAP projection of employee embeddings, colored by reason for leaving.

7.4 Employment duration

When coloring the embeddings by tenure (i.e., number of days employed), we see a similar outcome: some trends, but no clearly separable clusters. This is expected. Reducing 256 dimensions to 2 for visualization inevitably compresses and obscures much of the encoded structure. A lack of visible separation in 2D does not imply the model has failed to learn these features.

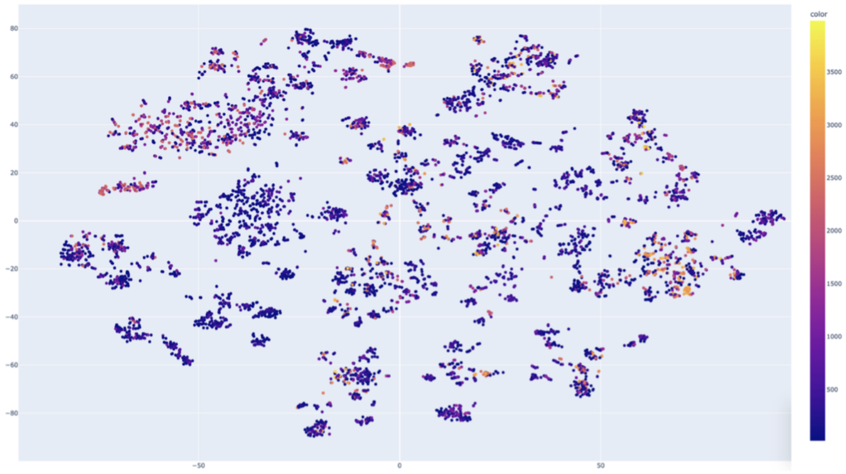

Figure 4: UMAP projection of employee embeddings, colored by employment duration.

7.5 What this tells us

These visualizations illustrate that the model has learned to embed employees based on meaningful behavioral patterns. Some attributes, like country, manifest clearly in the embedding space, while others such as tenure and exit reason are likely encoded in more complex, nonlinear combinations of dimensions.

To better surface these latent signals, we now move beyond visualization and leverage the pretrained model more directly. Specifically, we fine-tune it to predict:

- Time remaining until an employee leaves the company
- Reason for leaving, if applicable

By using the rich behavioral understanding the model has already acquired, this fine-tuning process becomes both more efficient and more accurate, allowing us to answer targeted questions without retraining from scratch.

8 Fine-tuning on attrition prediction

Having pretrained our foundation model to understand employee behavior through next-event prediction, we now fine-tune it on two downstream tasks of direct organizational relevance:

- Predicting why an employee will leave (including the case where they don't leave)
- Predicting when an employee will leave, expressed as the number of weeks into the future

These tasks are modeled sequentially at every point in each employee's behavioral history. This means that from the very first event to the most recent, the model is asked:

- Given all past behavior, how much longer is this employee likely to stay?
- If they leave, what is the most likely reason?

This design evaluates the model's predictive ability over the full timeline, rather than only at a single endpoint.

8.1 Training and generalization

In Figure 5, we show the loss curves for both training and test data. As the figure illustrates, loss converges smoothly, with test performance closely tracking training performance. This is strong evidence that the model generalizes well and is not overfitting.

Losses are decomposed into two components:

- Cross-entropy loss for predicting reason for leaving
- Cross-entropy loss for predicting time-to-exit (discretized into weeks)

Time-to-exit is dynamic: at each point in the sequence, the model must estimate how many additional weeks an employee will remain. Naturally, prediction improves as more data becomes available later in the sequence, but the model shows predictive power even early on.

8.2 Performance on a held-out test set

While cross-entropy is a robust metric, it's less intuitive than more familiar accuracy-based metrics. To provide a clearer picture, we constructed an evaluation using a balanced and truncated test set:

- Test size: 6,705 employees
- Class balance: 4,653 employees who will leave within the next month, 2,052 who will stay at least another month
- For each employee, a few months of recent behavior were randomly truncated, simulating a real-world partial observation scenario.

The model achieved:

- 87% true positive rate (4,046 / 4,653 employees correctly predicted to leave)
- 99.3% true negative rate (2,038 / 2,052 correctly predicted to stay)
- Overall accuracy: 91%
- Precision: 0.997
- Recall: 0.87
- F1 Score: 0.93

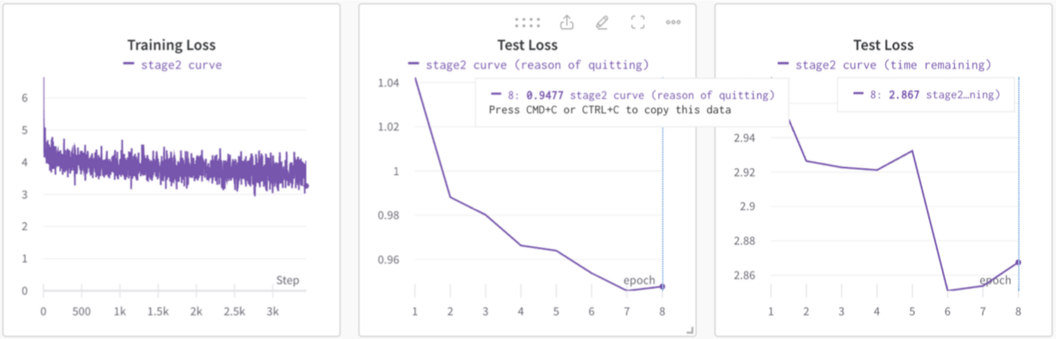

Figure 5: Cross-entropy loss during fine-tuning for attrition prediction.

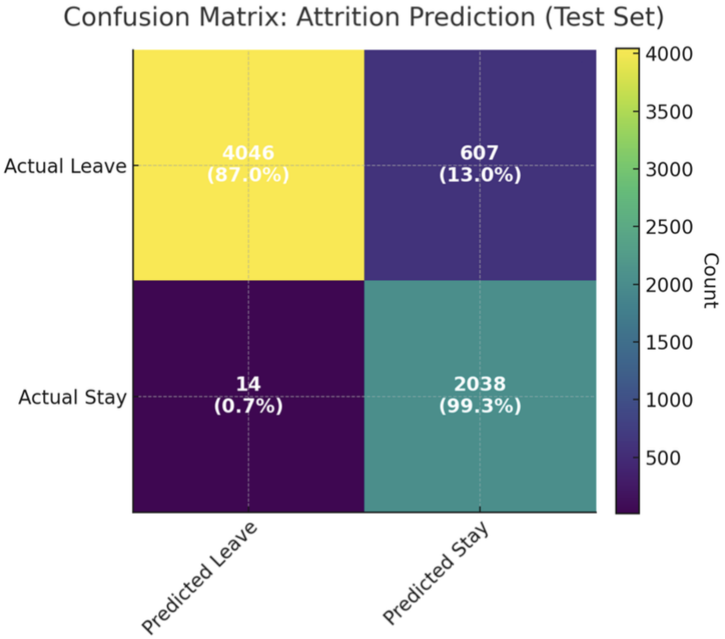

Figure 6: Confusion matrix of attrition predictions on held-out test data.

8.3 Real-world validation

To test the model in a high-stakes blind evaluation, one of our enterprise partners provided anonymized behavioral sequences for a group of employees, without revealing any information about their attrition. We were asked to use the model to predict attrition risk and identify the 10 most likely cases.

Result: 9 of the 10 had indeed left the company. The remaining individual was on an indefinite leave of absence.

This real-world challenge reinforced that the model's behavioral foundation generalizes, providing actionable insights even in ambiguous or partially observed scenarios.

9 How much does pretraining help?

A natural question is: Why bother with pretraining at all? Why not simply train a model directly on the downstream tasks of predicting when and why an employee will leave, using the raw behavioral sequences?

The answer lies in the power of generalization.

By first solving the harder, self-supervised task of predicting every action based on prior actions, the model learns to deeply understand the structure of employee behavior. This “behavioral language” pretraining builds a rich internal representation of employees’ routines, deviations, and patterns over time. I.e. you want to get to know them and use that knowledge as the foundation for understanding and predicting specific behaviors. Simplifying will remove key components of the underlying factors around who they are.

9.1 Quantitative impact

Pretraining led to a 7% improvement in test set performance on attrition prediction compared to training the same model architecture from scratch. While 7% might seem modest at first glance, it represents more than just a single accuracy metric:

- The pretrained model generalizes better to unseen employees, new roles, and future time periods

- It is less sensitive to data sparsity and requires fewer labeled examples of a specific task to perform well on that task

- It enables faster and more stable fine-tuning, making retraining less urgent and cheaper when new data arrives

9.2 Strategic benefit

Most importantly, pretraining lays the foundation for transfer learning across multiple tasks. Every improvement to the pretrained model boosts performance across all downstream use cases. I.e., not just attrition, but also productivity modeling, engagement prediction, and task allocation.

This is the essence of synergistic intelligence: one model, trained broadly, improving performance everywhere. Also, it scales—maintaining hundreds of specific downstream models and improving each one continuously is untenable, but always improving and optimizing one pretrained model is straightforward.

10 Why traditional data analysis didn’t work

Prior efforts to understand and predict employee attrition relied heavily on survey-based data, most notably, a dataset of approximately 10,000 employee responses across multiple self-reported metrics such as job satisfaction, engagement, and intent to stay.

In Figure 7, we show one such example: the correlation between self-reported satisfaction and subsequent days of continued employment. While there is a slight negative trend where employees reporting lower satisfaction tended to leave somewhat sooner, the signal is weak and noisy. In fact, domain experts designed these multiple-questions surveys and defined a score with the aim for it to positively correlate with tenure, but instead it proved to have a slight negative correlation, see Figure 8.

Despite extensive effort, including segmenting, reweighting, random forests, and model tuning, neither we nor earlier teams were able to produce a generalized or robust predictor of attrition from this data. Most results reflected top-down assumptions (such that reported satisfaction is a strong predictor of attrition) more than bottom-up insight, and became an exercise in wishful interpretation, rather than reproducible intelligence.

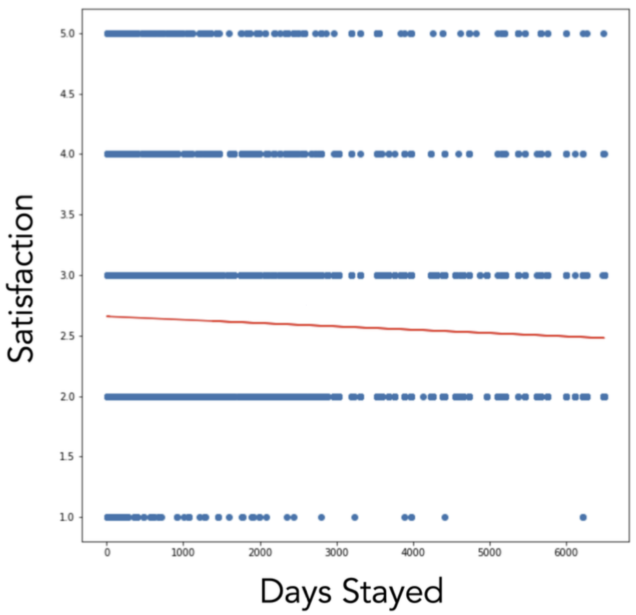

Figure 7: Relationship between self-reported satisfaction and continued employment. Red line shows average trend.

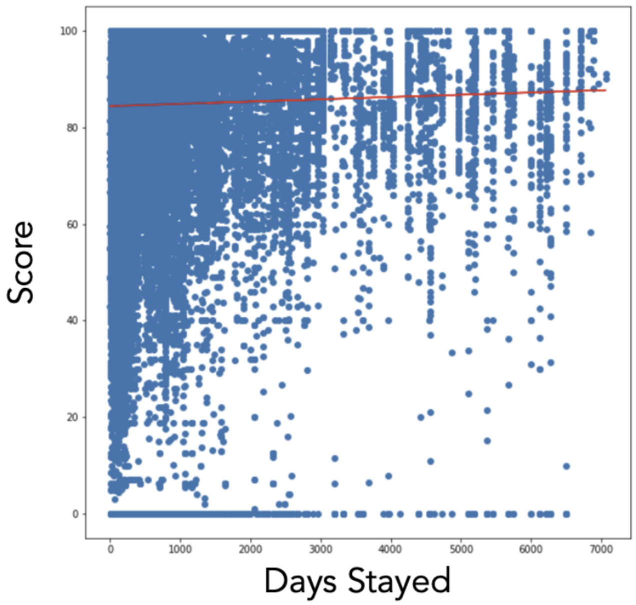

Figure 8: Relationship between expert-crafted score and continued employment. Red line shows average trend.

10.1 From analysis to modeling

This project highlights a broader shift: where traditional data analysis fails, Large Behavioral Models succeed.

Rather than relying on static reports or survey responses, our approach treats behavior itself as the data source—modeled as sequences, learned as a language. We do not need to ask employees what they think; we infer what they do. And that makes all the difference.

The future of AI for behavior-driven businesses is not in having LLM agents perform data analysis, but in having large models trained on behavioral data that embed rich understanding of people and processes. These models can then be fine-tuned and interpreted to solve complex, multi-dimensional challenges like attrition, productivity, or burnout, which are challenges that lie beyond the reach of dashboards and descriptive statistics.

In short: only large, expressive models, like Transformers with millions of parameters, are capable of compressing nuanced behavioral signals into actionable insight. Data analysis alone cannot get to know your employees. Foundation models can.

11 Business impact

Employee attrition isn't just a human capital issue, it's also a major financial liability.

Consider a company with 30,000 employees and an annual turnover rate of 70%. Even using a conservative cost-per-hire estimate of $5,000 (SHRM article), this translates to:

Annual attrition cost:
$0.7 \times 30{,}000 \times \$5{,}000 = \$105$ million per year

Now assume that a predictive system, like the one enabled by our foundation model, helps managers proactively retain just 15% of the at-risk workforce. That alone would result in $16 million in annual savings.

However, most studies estimate that the true cost of turnover (including recruiting, onboarding, lost productivity, and time to full performance) is closer to 50% of the employee's annual salary (G&A Partners article). At an average cost of $20,000 per departure (Qualtrics study), the actual annual cost of attrition rises to:

Annual attrition cost:
$0.7 \times 30{,}000 \times \$20{,}000 = \$420$ million per year

A 15% reduction in attrition would then yield $63 million in annual savings.

And that's just from attrition alone. When you add in burnout risk, underutilized talent, and reactive scheduling, the costs—and missed opportunities—grow exponentially.

Indeed, foundation models trained on behavioral data open the door to a much broader range of high-impact applications: identifying burnout risk, optimizing workforce allocation, improving engagement, and tailoring interventions at scale. The ability to truly understand employees based on their behavior (not just what they say, but what they do) offers enduring strategic advantage. This is just the beginning.

12 Conclusions

This work demonstrates that foundation model technology, powered by self-supervised causal modeling, can effectively capture and predict workforce behavior and dynamics. It achieves what traditional analytics and survey-based approaches have consistently failed to deliver.

By treating behavioral data as a language that is structured, sequential, and learnable, we showed that employee actions can be compressed, modeled, and predicted. Our model achieved 91% accuracy in forecasting whether an employee would leave in the following month, with high true positive and true negative rates. This is a significant result in a domain where predictive success has long eluded both statistical models and expert-driven analysis.

Importantly, this success was not the result of clever feature engineering, better dashboards, or more refined survey instruments. Instead, it came from training a large neural network to observe, learn, and internalize the behavioral patterns of employees, one action at a time.

The guiding principle of our approach is simple but powerful: Predicting what someone will do, means understanding who they are.

This philosophy moves beyond static metrics and reported satisfaction scores. It embraces the richness and nuance of actual behavior as the most reliable signal for understanding workforce dynamics.

As foundation models continue to evolve, their application to human behavior—within organizations and beyond—will redefine how we understand, support, and empower people at scale.

Citation

```
@article{unbox2025behaviorgptatwork,
  author  = {Rickard Br{\"u}el Gabrielsson and Vasudev Gupta and
      others},
  title   = {BehaviorGPT at Work: A Foundation Model for Workforce
      Actions \& Dynamics through Large Behavioral Modeling},
  journal = {Unbox AI Blog},
  year    = {2025},
  month   = jun,
  url     = {https://research.unboxai.com/behaviorgpt-foundation-model
      -workforce.html}
}
```