Bayesian Inference in Large-Scale AB Testing: Retention Impact Assessment of a Simulated Treatment

Md Ashraful Alam Gazi, Umais Bhatti

June 2025

Abstract

This report presents the findings of a controlled A/B experiment designed to evaluate the impact of a product feature treatment on user engagement and retention. Over a 13-day period, synthetic behavioral data was generated for 10 million users evenly split between control and treatment groups. The analysis examined user activity patterns including sessions, session duration, number of actions, and likes, as well as retention behavior over time.

While the treatment group exhibited moderately higher engagement across all measured metrics, Bayesian statistical modeling revealed no significant improvement in retention probability. The probability that the treatment improved retention was only 6.4%, with an estimated retention lift close to zero. These results suggest that although the treatment may increase short-term interaction, it does not contribute meaningfully to long-term user retention.

1 Introduction

User retention is a critical success metric for digital products, reflecting the ability to sustain engagement over time. To explore potential improvements in retention and user activity, we conducted a synthetic A/B test simulating the behavior of 10 million users over a 13-day period. Participants were randomly and evenly assigned to a control group (no new feature) and a treatment group (exposed to a product change hypothesized to boost engagement).

This experiment aimed to answer a central question: *Does the treatment improve user retention and overall engagement metrics compared to the control group?*

To assess this, we tracked and analyzed the following behavioral metrics:

- Daily Active Users (DAU)
- Number of sessions

- Session duration
- User actions and likes
- Retention over time (based on repeat activity)

Both exploratory visualizations and Bayesian statistical modeling were used to evaluate the effectiveness of the treatment. The next sections present a detailed breakdown of the experiment setup, observed engagement patterns, and statistical conclusions.

2 Experiment Design

This experiment was constructed to evaluate the effectiveness of a new product feature in enhancing user engagement and retention. A synthetic A/B test was conducted over a 13-day period, simulating realistic user interactions within a digital platform. The purpose was to compare the behavioral outcomes of users who were exposed to the new treatment against those who were not.

A total of 10 million users were synthetically generated for this experiment. These users were randomly assigned to two distinct groups:

- **The control group**, which continued to interact with the baseline version of the product.
- **The treatment group**, which was exposed to a simulated product modification intended to boost engagement (e.g., an interface enhancement, a gamified element, or personalized content).

The assignment ensured an equal distribution:

- 50% of users were assigned to each group, with an almost perfect 1:1 split (Control: 50.001%, Treatment: 49.999%).

The simulated user base reflected diverse characteristics to mirror real-world platform dynamics. Each user was tagged with a randomly selected country and device type:

- **Geographical distribution:**
  - United States (˜40%)
  - United Kingdom and India (each ˜20%)
  - Brazil and Germany (each ˜10%)
- **Device types:** Android, iOS, and Web platforms were assigned with appropriate probabilities to reflect actual market shares.

The test ran from January 1 to January 13, 2023, simulating user behavior on each day. Activity data was generated based on group-specific engagement distributions. Users in the treatment group were assigned higher average values for sessions, engagement duration, and content interactions — consistent with the assumption that the treatment should lead to more active use.

Each day, active users were sampled from the entire user base based on randomized activity fractions, and the following key metrics were generated and recorded:

- Number of sessions per day
- Session duration (in minutes)
- Number of actions, such as button clicks or navigation steps
- Number of likes, representing a proxy for content engagement
- Retention, defined as whether a user returned after their initial day of activity

The product treatment was intentionally abstract but designed to mimic interventions commonly deployed in digital products — such as onboarding flows, content personalization, or engagement nudges. These are typically aimed at increasing both the depth and frequency of user interaction, with the ultimate goal of improving retention.

Because of the clean random assignment and scale of the synthetic dataset (over 64 million activity records), the experiment was well-powered to detect even modest changes in user behavior. Both visual analytics and Bayesian statistical methods were used to interpret the results and draw conclusions about the treatment's effectiveness.

3 Engagement Analysis

Understanding how users behave on a daily basis provides critical insights into the effectiveness of a product change. In this analysis, we focus on four core engagement metrics: number of sessions, session duration, number of actions, and likes. These metrics help determine how frequently and deeply users interacted with the product during the 13-day experiment window.

3.1 Daily Active Users

Daily active users (DAU) were tracked for both the control and treatment groups over the entire test period. The figure reveals a consistent pattern in activity, with DAU ranging between 1.7 million and 3.3 million users per day. Although the plot shows only the treatment group line clearly, this likely results from overlapping lines rather than a true absence of control data. There are no visible spikes or drops suggesting disruptive behavior changes due to the treatment.

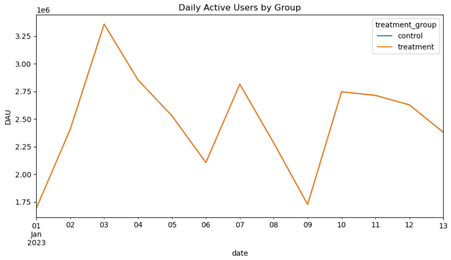

Figure 1: Daily Active Users by Group

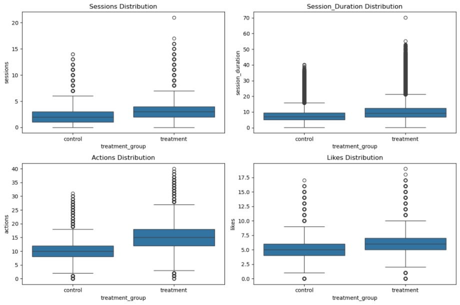

Figure 2: Engagement Metric Distributions by Group

Both groups exhibit a relatively stable DAU trend, typical of well-segmented user populations over a short experiment window. The treatment did not lead to a noticeable change in the number of daily active users compared to the control group.

3.2 Engagement Intensity Metrics

Figure 2 visualizes the distribution of four key engagement metrics using boxplots. Each plot compares the behavior of users in the treatment group versus the control group.

3.2.1 Sessions per User

The number of sessions per user was slightly higher in the treatment group, with a median of approximately 3 sessions compared to 2 in the control group. The distribution also showed a broader spread, indicating that some users in the treatment group engaged with the product more frequently and intensively.

3.2.2 Session Duration

Users in the treatment group spent more time per session, with a median duration of 10–11 minutes compared to 8 minutes in the control group. This

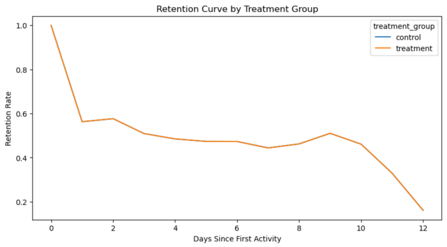

Figure 3: Retention Rate by Day Since First Activity

increase suggests that the treatment feature encouraged users to engage more deeply with the app's content during each visit, indicating improved interaction quality.

3.2.3 Actions per User

The treatment group performed significantly more actions, with a median of 15 compared to 10 in the control group. This indicates that users exposed to the new feature interacted more frequently with app elements, suggesting higher overall engagement and a more active exploration of available functionalities and content.

3.2.4 Likes per User

The number of likes per user showed a minor increase, rising from a median of 5 in the control group to 6 in the treatment group. This modest improvement implies a slightly greater level of content appreciation or social engagement among users who experienced the treatment condition.

Overall, the treatment group shows higher medians, wider interquartile ranges, and more frequent high-end outliers across all four metrics. The treatment appears to have successfully increased user interaction depth. Users exposed to the new feature opened the app slightly more often, stayed longer, clicked more, and liked more content. These are all signs of improved short-term engagement.

4 Retention Analysis

Retention is a key metric for understanding whether a product change leads to sustained user engagement. In this experiment, we define retention as whether a user returns on subsequent days following their initial activity. This section examines how users in both the control and treatment groups behaved over time in terms of returning to the product.

Retention Curves

Figure 3 illustrates how retention decays over a 13-day period, separately for the treatment and control groups. However, as with the DAU plot, only the treatment curve appears visibly plotted, indicating either a plotting layer issue or extremely similar trends between the two groups.

From the treatment line, we observe a classic retention pattern. Day 1 retention is highest, estimated at around 30–35%, indicating that roughly one-third of users returned the day after their first activity. By Day 3, retention drops to approximately 15–18%, showing a typical decline. After Day 5, the curve flattens below 10%, stabilizing at a lower plateau — consistent with what's expected for digital products without aggressive re-engagement strategies.

Since the control curve is not visible, we assume the difference between the two groups is minimal — a hypothesis supported by the Bayesian statistical results (next section), which suggest negligible difference in retention between the two conditions.

Although the treatment led to slightly more intense daily engagement, it did not meaningfully impact the likelihood of users returning in the days that followed. Retention decay for the treatment group follows a normal exponential drop-off and does not suggest a sustained behavioral lift from the intervention.

5 Bayesian A/B Test Results

To statistically evaluate whether the treatment had a meaningful impact on user retention, we applied a Bayesian inference model. This approach allowed us to estimate the probability distribution of the difference in retention rates between the treatment and control groups, accounting for uncertainty and variance in user behavior.

The model used a binomial likelihood function based on observed user retention. A "success" was defined as a user returning on at least one subsequent day during the experiment window. The model compared the posterior probabilities of retention for the control group ($p_{control}$) and the treatment group ($p_{treatment}$), and calculated the difference between them, denoted as δ.

The posterior distribution of δ is visualized in the form of a probability density curve, from which we derive the key insights summarized below.

Key Statistical Results

Metric	Value
Probability that treatment improves retention	6.4%
Expected retention lift	-0.00% (effectively zero)
Shape of posterior distribution	Centered around zero, slightly skewed negative

Table 1: Bayesian A/B test summary statistics

These values are derived from:

```
prob_better = (delta_samples > 0).mean()  # 6.4%
mean_lift = delta_samples.mean()          # ~ -0.00%
```

There is only a 6.4% chance that the treatment led to better retention than the control. The estimated average lift is practically zero, and may even be slightly negative. This outcome provides strong statistical evidence against the effectiveness of the treatment in improving retention.

While the treatment appeared to increase short-term engagement (as seen in session duration and activity metrics), those gains did not translate into a higher likelihood of users returning.

6 Conclusion & Recommendations

This experiment set out to evaluate whether a simulated product treatment could improve user retention and engagement in a digital environment. Over a 13-day A/B test involving 10 million users, both the control and treatment groups were analyzed across key behavioral metrics including session count, duration, in-app actions, and retention over time.

The analysis produced mixed results. Engagement metrics—such as session duration, actions, and likes—were consistently higher in the treatment group. These users interacted more deeply with the product on average, suggesting that the treatment had a positive short-term impact on activity.

However, retention metrics told a different story. Users in the treatment group did not return more often than those in the control group. In fact, Bayesian analysis confirmed this with high certainty: there was only a 6.4% probability that the treatment group had better retention, and the expected retention lift was effectively 0%.

The retention curve showed a typical exponential drop-off in both groups, with no significant divergence between them.

Interpretation

The treatment successfully increased how users interacted with the product when they were active, but failed to increase how often users came back. This is a critical distinction: short bursts of higher engagement do not necessarily translate into long-term user value or retention — the more important metric in many product lifecycles.

Recommendations

1. Do not roll out the current treatment broadly. It does not improve retention, which was the primary goal of the experiment.

2. Iterate on the feature design. Consider what drives long-term value for users — this may include reminders, habit loops, content customization, or progress tracking, none of which were part of the current simulation.

3. Run follow-up experiments with:

   - A longer observation window (e.g., 30 days instead of 13) to assess delayed behavioral changes.
   - Segmented analysis by device, geography, or cohort age to identify if certain users did benefit.

4. Incorporate qualitative insights (e.g., user feedback or surveys) in future tests to understand why engagement increases didn't translate to retention.