Can simulated shoppers predict A/B tests? The Squoosh Bench preprint explained

22 November 2026 · 6 min read · blogs.iictp.org/squoosh-ux-1-simulated-shoppers-ab-test-forecast

Short answer: Squoosh Bench is an arXiv preprint dated 18 November 2026 that asks whether forecasts from simulated shoppers can help decide which website changes are worth a live A/B test. The authors report that Squoosh UX-1, a model fine-tuned on human browsing data, classified 267 of 318 historical test outcomes correctly (84.0%), against 61.6% for the strongest general-purpose model they tried. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- Squoosh Bench, an arXiv preprint dated 18 November 2026, tests whether simulated shoppers can forecast which website changes win A/B tests.
- The authors report that UX-1 classified 267 of 318 historical experiments correctly (84.0%), against 61.6% for Qwen3 Next 80B.
- It is an unreviewed preprint on a screened set of tests, so use such forecasts only to triage your testing backlog and back-test them on your own experiments.

Read online, listen to it, or change the text size: https://blogs.iictp.org/squoosh-ux-1-simulated-shoppers-ab-test-forecast

What is Squoosh Bench?

Experimentation teams must decide which proposed changes to spend live traffic on. The paper tests a prediction step before that: simulate shoppers, forecast the result, and use the forecast to prioritise. The authors screened 412 historical website A/B tests and evaluated four systems on the 318 that ran longer than 21 days. Each system predicted an original win, a variant win or an inconclusive result.

How it works

- Page roles were masked and recorded outcomes were withheld during inference.
- Squoosh UX-1 simulated 1,000 shoppers per arm of each test.
- The comparison model was Qwen3 Next 80B, described as the strongest general-purpose configuration on outcome accuracy.
- The score is classification accuracy over three outcomes, original wins, variant wins or inconclusive.

Results reported by the authors

- UX-1 classified 267 of 318 correctly: 84.0% with a 95% Wilson interval of 79.5–87.6%, treating experiments as independent.
- Qwen3 Next 80B classified 196 of 318 correctly: 61.6%.
- UX-1 favoured the variant in 126 experiments; in the live tests 118 of those variants won (93.7%), four lost and four were inconclusive.
- The remaining 192 experiments included 19 additional variant wins.

Limitations and open questions

- It is a preprint. UX-1 is the headline system named in the title; whether the authors built it, or are affiliated with its maker, was not checked here.
- The tests are historical and from the authors' screened set, so results may not carry to your site, traffic mix or product.
- The 192 experiments without a UX-1 variant call contained 19 variant wins, so the system does not find every winner. A forecast prioritises tests; it does not replace running them.
- The wording "treating experiments as independent" is the authors' caveat on the interval.

How to apply the idea carefully

1. Treat any simulated-shopper forecast as a triage score for your testing backlog, not as a result.
2. Back-test on your own closed experiments with outcomes hidden, as the paper does.
3. Track false positives and false negatives separately, since a missed winner and a wasted test cost differently.
4. Keep running live tests for changes with high revenue or risk.

What it means for marketers

If the approach holds on your data it could lower the cost of idea screening. The key discipline is the one the paper models: hide outcomes, define the classes in advance and report intervals. See how to evaluate an LLM application.

Frequently asked questions

- **Does this replace A/B testing?** No. The paper's use is deciding which changes to test live.
- **Is 84% accuracy good?** It is the authors' reported figure on their test set; compare it with a baseline on your own history.
- **Is it peer reviewed?** No, it is a preprint.

Related reading

- How to evaluate an LLM application
- What is LLM hallucination?
- AI Practitioner

Sources and currency

Sources were read on 22 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Pasula R., Dramiga J., Shah V., Halim S., Dooskin T. Squoosh Bench: Evaluating UX-1 on 318 Historical Website A/B Tests. arXiv:2611.25200. 18 November 2026. arxiv.org/abs/2611.25200.

Published by IICTP, an independent technical credential body, on 22 November 2026. The current version of this post is at https://blogs.iictp.org/squoosh-ux-1-simulated-shoppers-ab-test-forecast. More guides: blogs.iictp.org.