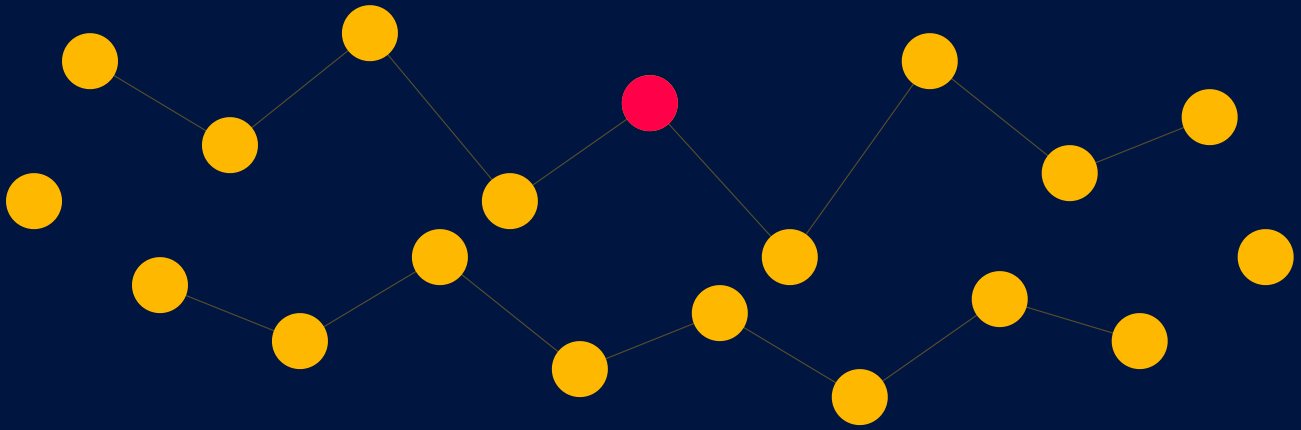




How AI is changing **who builds** experiments and how.

Findings from nineteen long-form interviews with senior experimentation practitioners on the current state and near future of prompt-based A/B testing.

JONNY LONGDEN Chief Growth Officer, Speero

RESEARCH SPONSORED BY

 **Kameleoon**

speero.com/research

INSIDE THIS REPORT

What's in this **paper.**

A short read designed to be scanned in fifteen minutes. Deeper detail lives in the full LinkedIn article.

01	Executive summary	03
02	What this research was and was not	04
03	How PBX works, in plain terms	05
04	The benefits practitioners actually care about	06
05	What unlocks this: context	08
06	Honest limitations	10
07	A watch-out: rigour does not come from the build tool	12
08	Where this tech is going	13
09	So, what is the answer?	14

ABOUT SPEERO

Speero is an experimentation agency and consultancy. We help companies design, run and learn from experiments that drive real growth, not just move metrics.

ABOUT KAMELEOON

Kameleoon is the only prompt-based experimentation platform that enables any team to turn ideas into live A/B tests in minutes, simply by chatting with AI. Built for modern growth teams, Kameleoon adapts to any stack and skill level, offering accurate data, advanced privacy, and native integrations with leading tools. More than 1,000 brands, including Toyota, Mayo Clinic, and Lululemon, trust Kameleoon to scale experimentation without compromising control.

01 • EXECUTIVE SUMMARY

The right question is no longer whether AI **replaces developers.**

AI variation building is the most credible change to happen to the A/B testing tool category in a decade. Every practitioner in this research had a visible first-impression reaction the moment they watched a sentence of plain English render as a live change on a real page.

The original framing of the research was *"can AI replace developers in A/B testing?"*. After nineteen long-form interviews, that turned out to be the wrong question. The more accurate one, and the one this paper answers, is how AI is changing **who gets to build a test** and what work is important for the people who used to do it.

RESEARCH AT A GLANCE

INTERVIEWED

19

Senior experimentation practitioners

COVERING

6

Industries · retail, SaaS, travel, fintech, education, healthcare

FORMAT

2h

Long-form video sessions, hands-on with PBX

INCLUDING

1

Kameleoon customer with 7 to 8 months of production use

AI variation tools meaningfully expand who can build experiments. They cut the bottleneck that has meant only the best-resourced or most politically connected tests get run. They unlock the category of test that would otherwise be cut from a quarterly roadmap. Engineers shift toward more complex work that genuinely needs them. Variations that win move cleanly into production through capabilities like Kameleoon's PBX Ship.

The bigger question this raises for experimentation as a discipline is about what programs do with the bottleneck that has just been removed, and who gets to participate in the discipline that it used to gatekeep.

02 • WHAT THIS RESEARCH WAS AND WAS NOT

Nineteen senior practitioners. First-impression evaluations, one power-user case.

Most of the nineteen practitioner sessions were structured as initial-impressions evaluations rather than power-user deep dives. Each practitioner spent up to two hours with PBX, built variations on their own sites, prompted it, refined the output, and discussed what they saw.

One of the nineteen, Chris Strobl at Cedars-Sinai, is an existing Kameleoon customer who has been using PBX in production for seven to eight months, including PBX 2.0. His observations are flagged where they come from sustained use rather than first impressions.

Kameleoon released PBX 2.0 during the research window. The practitioners and I worked with the earlier release. This paper flags where 2.0 changes the picture.

SIX INDUSTRIES

Sectors represented in the research

Retail	●
SaaS / Tech	●
Travel	●
Fintech	●
Education	●
Healthcare	●

CRO LEADS • PMs • DEVS • DATA SCIENTISTS

FIG. 01 • SECTORS REPRESENTED

TRANSPARENCY

This is sponsored research.

Kameleoon paid for the research and provided access to the PBX tool. They asked for an honest read, not a sales piece, and did not intervene in the findings. Where PBX-specific capabilities are described, the paper is explicit. Where category-level observations are made, they apply to all AI variation tools.

03 · HOW PBX WORKS, IN PLAIN TERMS

A browser extension that **reads the live site.**

PBX runs as a browser extension. Open a site, type a prompt describing what you want changed, and PBX gathers page context automatically (DOM, CSS, structure, screenshots). It then generates and applies the JavaScript and CSS needed to preview the variant live, and gives you the code.

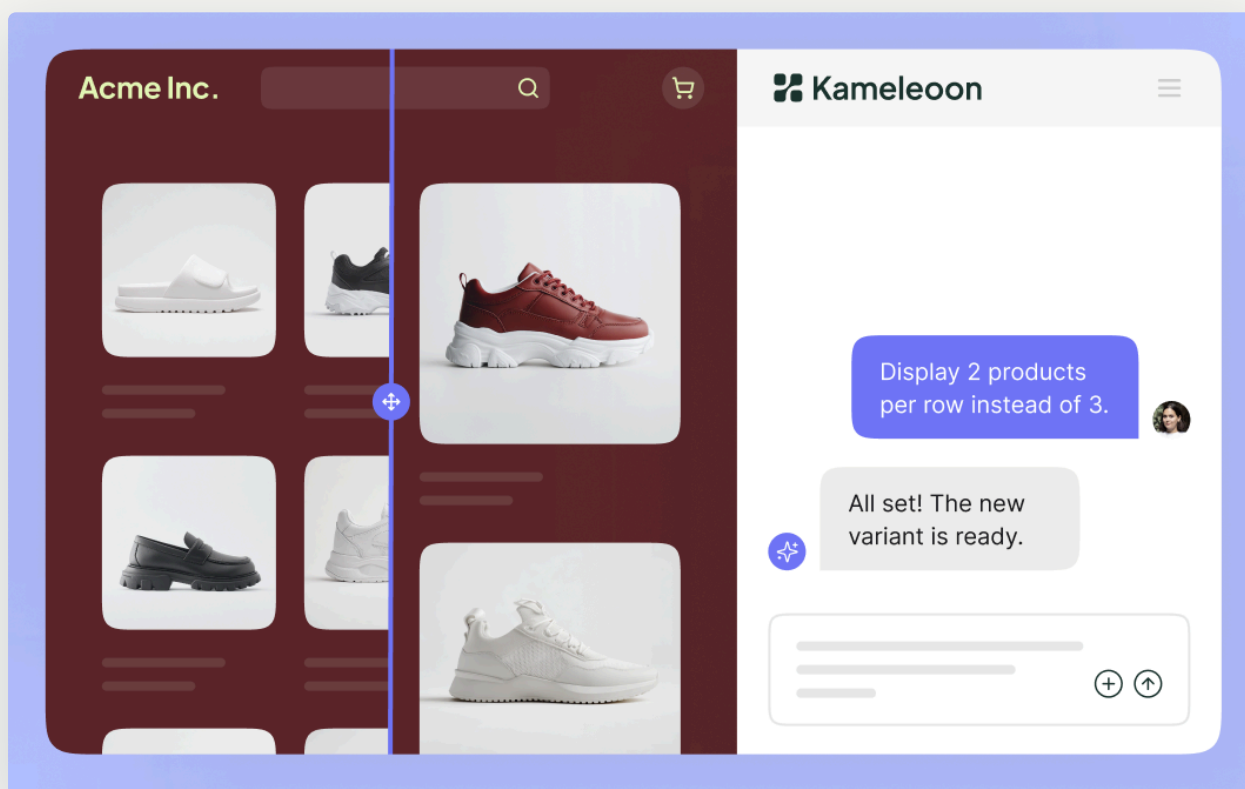


FIG. 02 · PBX 2.0 RENDERING A VARIANT ON THE LIVE SITE FROM A PLAIN-ENGLISH PROMPT

The variant is not a separate mockup built in a different environment. It is a working variant of the actual page, built inside the actual context of your actual site. From there, the same variant can be ramped through Kameleoon's platform, complete with feature flagging, audience targeting, and event tracking.

A tool like Claude or Lovable can build a static prototype that looks like your site. But the prototype lives somewhere else and has to be rebuilt to become a test. That is the structural difference that makes the category interesting.

FROM IDEA TO LIVE TEST

Days or weeks become minutes.

From prompt to live test in one continuous artefact

The variant does not have to be rebuilt to become a test.

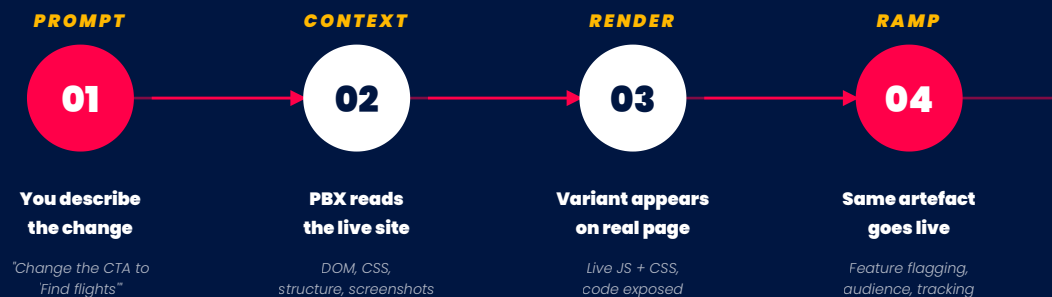


FIG. 03 · THE PBX WORKFLOW

WHY THIS MATTERS

The variant is the test, not a mockup of the test.

This is the structural difference between PBX and general-purpose AI coding tools. A prototype built elsewhere has to be rebuilt to become a live experiment. PBX collapses those steps into one continuous artefact.

04 • THE BENEFITS PRACTITIONERS ACTUALLY CARE ABOUT

Not what the vendor narrative expects.

The benefits practitioners flagged were not raw speed and cost savings. Those came up. The more interesting answers were about **access**.

— THE DEV QUEUE HAS GOT WORSE, NOT BETTER

A consistent theme across every interviewee: getting developer time to build a test is harder now than three years ago, not easier. Every interviewee described the same dynamic, including practitioners at enterprises with significant engineering headcount. Developer bandwidth is constrained everywhere.

"WGU went through a return-to-office mandate. Development resources have become very scarce. To have a test go through the development queue takes quite a long time. Wherever possible I build what I can. If I have to go through dev it can take up to eight weeks."

WHITNEY NORTON • TESTING AND OPTIMIZATION MANAGER • WESTERN GOVERNORS UNIVERSITY

— WHICH TESTS ACTUALLY GET RUN

Every CRO lead I spoke to has a backlog of test ideas that are interesting but not interesting enough to justify a dev sprint. Those tests almost never get run. AI variation tools quietly change that calculation. The cost of trying an idea drops below the threshold at which it needs to be defended.

"I've had many ideas over the years that no one was going to get excited about. Usually I'd need to build up credibility and trust first to convince stakeholders to take on these unique ideas. Now I can do a lot of these tests myself. I don't even have to pitch them half the time."

BEN YOUNG • CRO AT HUBSPOT

— REALLOCATION, NOT REPLACEMENT

This is the shape of the shift. Engineers move toward the complex work that genuinely needs them. Practitioners absorb the variation building that should never have required engineering attention in the first place.

"In the past, all of the coding was done by the development team. Now, I do most of the coding myself because of AI. The developers only handle the difficult tests."

BEN YOUNG • CRO AT HUBSPOT

— STAKEHOLDER BUY-IN IS WHERE THE IMMEDIATE VALUE LIVES

Practitioners independently described the same use case: showing a stakeholder an idea in a meeting, on their real site, in a state they can interact with. The mockup that wins the room is the same artefact that can be pushed into a live test.

"Having something that you could play with in real time and just say, this is a working project, this is kind of what we're thinking, something actually baked into the site rather than you trying to hack something together. I see that as being really useful."

FAITH DALLAS • DIGITAL EXPERIMENTATION MANAGER • NEW LOOK

THE FINDING WORTH CARRYING

The bottleneck removed is not what most people assume.

Practitioners consistently talked about access, not speed. The mechanical build step getting faster is the surface benefit. The deeper shift is which ideas get to be tests at all, and which people in the organisation can participate in experimentation.

05 • WHAT UNLOCKS THIS: CONTEXT

The most defensible reason to use PBX over Claude Code.

Multiple practitioners raised, directly or by implication, the question that will follow this category around: why would a developer use a specialist tool like PBX when they already have Claude Code, Cursor or GitHub Copilot for everything else? The answer that emerged from the research was one word: **context**.

CONTEXT • THE HIDDEN WORK

GENERAL AI CODING TOOL	PBX
<i>Claude Code · Cursor · Copilot</i>	<i>Runs against the live site</i>
DEVELOPER MUST SUPPLY	AUTOMATICALLY LOADED
→ Paste HTML → Paste relevant CSS → Describe the framework → Name the selectors → Explain the brand rules	→ Live DOM as rendered → Live CSS in effect → Framework and structure → Selectors as they exist → Brand rules via master prompt (2.0)
Repeated every test.	Loaded once. Reused forever.

FIG. 04 • GENERAL AI CODING TOOL VS PBX • THE CONTEXT GAP

"What PBX is trying to do is what we found missing in other tools. The context."

SENIOR OPTIMISATION MANAGER • DEVELOPER BACKGROUND • MAJOR AIRLINE

— PBX 2.0 EXTENDS CONTEXT INTO INSTITUTIONAL TERRITORY

Kameleoon released PBX 2.0 during the research window. It adds two capabilities that push the tool further into contextual territory: connection to a brand's Figma design modes, and a master prompt where an organisation admin encodes design system rules, accessibility constraints, brand voice, and website-specific technical instructions. The tool moves from "faster AI build" toward "an AI build environment that already knows what your site is supposed to look like".

One customer using PBX 2.0 in production described the effect concretely:

"With PBX 2.0, it's got even more capabilities to where it's aware of more of your website, not just the single page. If I have a test I need to run on a dozen pages, it realises, oh, this seems to be kind of like a template. So, do you want me to run this test on any page that is of this template? I'm like, yes, please."

CHRIS STROBL • CEDARS-SINAI • SEVEN TO EIGHT MONTHS ON PBX IN PRODUCTION

THE CEILING

Not the model. What surrounds the model.

The most transferable insight from this research: the limit on what AI variation tools can do is less about the underlying language model and more about the contextual scaffolding around it. Tools that get design system, brand voice, accessibility rules and prior test learnings into the AI's default context will produce visibly better output.

06 · HONEST LIMITATIONS

Where practitioners **hesitated.**

— TRUST ON THE LIVE SITE

Practitioners want to inspect and QA any code that could touch live traffic, regardless of who wrote it. PBX exposes the code it produces, supports simulate mode, and allows small-percentage traffic ramps before full release. The build step gets faster. The release process does not change as much.

"If it spits out the code, I would probably put that into a lower environment and test it before putting it into production. I would be curious if the AI code affected something else that a more in-depth QA process would need to catch."

HOLLY GLEASON · DIGITAL TESTING MANAGER · US FOOD RETAILER

— SINGLE-PAGE APPLICATIONS

In the version practitioners tested, SPA frameworks were a limitation (the airline example was Angular). Kameleoon report that PBX generates code using their JavaScript API (`runWhenElementPresent`, `enableDynamicRefresh`) and that the admin master prompt can encode Shadow DOM handling and custom components for the site. A customer running PBX in production on a SPA site described the opposite experience:

"It works quite well, almost surprisingly well. With Target you basically have to bake something into your web pages to inform Target that the page changed. Whereas with Kameleoon, the automatic detection has been great."

CHRIS STROBL · CEDARS-SINAI · SUSTAINED PRODUCTION USE

— VALIDATION FOR NON-DEVELOPER USERS

A data scientist with experience on multiple enterprise experimentation platforms flagged the gap between prompting and debugging. Kameleoon's Configure and Simulate features in PBX 2.0 let users pre-validate test assignment and event firing before launch.

07 · A WATCH-OUT FOR THE DISCIPLINE

Rigour does not come from the build tool.

CRO and experimentation exist as disciplines partly because they impose rigour on what would otherwise be opinion-driven, political or whimsical product decisions. The rigour is not in the variation itself. It is in the hypothesis, the prior research, the success criteria agreed before the test runs, the analysis plan, the statistical thresholds, and the post-experiment write-up.

The build tool sits in the middle of this process. It does not produce the rigour around it.

"It's one of those dangerous things that could get out of hand if it's not actually governed."

SENIOR CRO LEAD · LARGE GLOBAL RETAILER

THE PRACTICAL TRANSLATION

Reinvest the time. Do not skip.

As AI absorbs variation-building work, make sure the work it gives back is being reinvested in the discipline that surrounds the variation. Better hypotheses, better instrumentation, better analysis. The tools are not designed to protect the rigour for you.

08 · WHERE THIS TECH IS GOING

The pipeline into production.

Prompt-based building is only the front end of the shift. What comes next is the pipeline that connects a winning variation to the codebase that actually ships.

Kameleon's PBX Ship extends the workflow into an engineer's IDE via an MCP server. Once a variant wins, the engineer can bring the same artefact into the production repository through PBX Ship, behind a feature flag. The prompt-built variation and the shipped feature are the same object at different stages of the same continuous workflow.

For the wider category, the direction of travel is multi-agent. Different agents handling different parts of the experimentation loop, coordinating through shared context. The ability of a specialist tool to hand off cleanly to production is what will separate the tools that keep pace from those that stall as one-off variation builders.

THE CATEGORY SIGNAL**Prompt to variation to production, in one continuous workflow.**

The AI variation tools that endure will be the ones that close the loop between the prompted variant and the shipped feature. That is a solvable problem, and PBX Ship is one credible answer to it already in the market.

09 • SO, WHAT IS THE ANSWER?

Three things AI variation tools actually change.

The right framing is augmentation, not replacement. Across the full range of practitioners, the consistent finding was that these tools change three things.

01

Who can build a test.

The practitioner with an idea is no longer dependent on engineering capacity, design availability and political bandwidth to find out whether the idea is worth pursuing.

02

Which tests get run.

Second-class ideas that used to die in the backlog now get a chance. The cost of trying one has dropped below the threshold at which it needs to be defended.

03

What specialists do.

The work absorbed is the variation-building that should never have required specialist attention. The work amplified is the strategic, judgment-heavy work.

"For the first time in the sixteen years I've been doing this, I'd be comfortable with more of a marketing type person going in there and doing the prompting."

CHRIS STROBL • CEDARS-SINAI • 16 YEARS IN EXPERIMENTATION

The bigger question for experimentation as a discipline is no longer whether AI replaces anyone. It is what programs do with the bottleneck that has just been removed, and who gets to participate in the discipline that the bottleneck used to gatekeep.

WHAT NEXT

The bigger question is what programs do with the bottleneck that has just been **removed**.

TRY THE TOOL

PBX is free to trial.

Install the browser extension and prompt your first variation on your own site.

Start your free trial →

kameleoon.com/pbx-free-trial

Full article on LinkedIn including practitioner biographies and additional context. [linkedin.com/pulse/...](https://www.linkedin.com/pulse/...) →