



SEPTEMBER 15, 2026

"NOBODY WAS FRUSTRATED"

The promise and problem of synthetic users in UX research

DFFRNT

DOMINIRA SAUL



SPEAKER NOTES

"I want to talk about an idea that is becoming increasingly plausible: doing user research without users."

Introduction

Dominira Saul

20 years in experience, human factors, user research, interaction design.

- President & Co-Founder, DFFRNT
- P/T Faculty, uOttawa, Digital Transform & Innovation
- Director UX, You.i TV
- Partner & Chief Experience, Akendi
- P/T Faculty, London Guildhall & Carleton
- Principal, usabilityagency.com
- MSc User Interface Design
- Phd Digital Transformation & Innovation

Company Confidential ©DFFRNT Inc. All Rights Reserved.

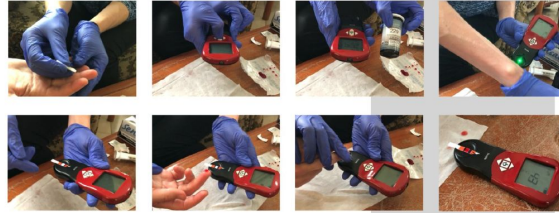


SPEAKER NOTES

Usability Testing TrueHB Hemoglobin meters with rural nurses in Ghana



Process for testing blood hemoglobin with the TrueHb Hemometer

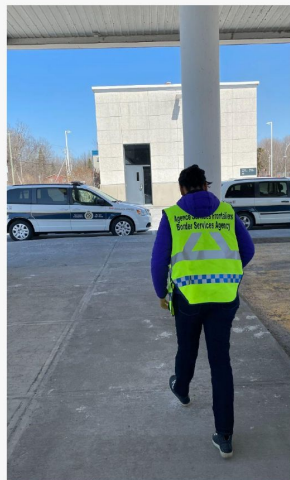


Note: for the purposes of our test we used simulated blood.

SPEAKER NOTES

Usability testing set-top box UI in Argentina & Brazil





SPEAKER NOTES

Shadowing medical surgical nurses in Milwaukee /
Observing travellers at various border crossings
(here is St. Stephen NB)

We have been incredibly fortunate to have had the opportunity to do user research in all parts of the world and involving all sorts of potential solutions whether that's Africa or South America all over Canada or the United States we have had the immense privilege of being able to interact with users in order to ascertain what is the best experience for them in a multitude of situations. this type of research is of course time-consuming and expensive. Is it a thing of the past?

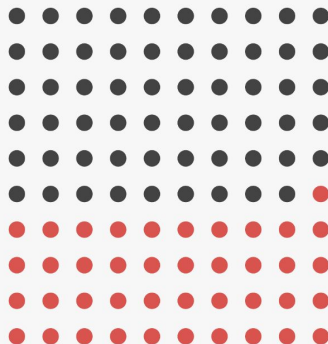
Having the option to have synthetic users provide you with real world, data or data approximating real world situations is tempting because of the efficiency it save time it save money it says effort, but the question is is it an actual replacement for first hand, boots on the ground, primary research?

The Seduction

"I interviewed 300 customers yesterday"

- They represented every age group, income bracket and level of technical expertise.
- I tested three designs, asked follow-up questions and segmented the results.
- The report was ready before breakfast.

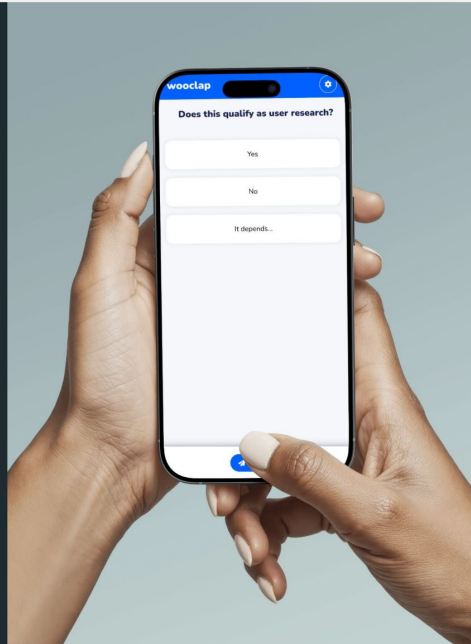
Company Confidential ©DIFFRINT Inc. All Rights Reserved.



SPEAKER NOTES

"Imagine I tell you I interviewed 300 customers yesterday. They represented every age group, income bracket and level of technical expertise. I tested three concepts, asked follow-ups, segmented the answers, and had the report ready before breakfast. Recruitment time: zero. Participant incentives: zero. There is just one catch... None of them exist."

In your opinion, does this qualify
as user research?



INSERT SLIDO

Definition

What is a synthetic user?

An AI-generated representation that produces responses or behaviours intended to stand in for those of a human user or user population.



PERSONA



RESPONDENT



PARTICIPANT



AGENT

Working synthesis informed by Hämmäläinen et al. (CHI '23), Salminen et al. (CHI '24), Xiang et al. (CHI '24), and Lu et al. (CHI EA '25).

Taxonomy

Four very different things are being called “synthetic users”

They make progressively stronger methodological claims.

1

Synthetic persona

Describes a hypothetical user

2

Grounded persona

Represents existing research/data

3

Synthetic participant

Answers research questions as a user

4

Synthetic agent

Attempts tasks or interacts with interfaces

Examples: Salminen et al. (CHI '24); Wang & Siu (CHI '26); Hämmäläinen et al. (CHI '23); Xiang et al. (CHI '24).

SPEAKER NOTES

Taxonomy (2 minutes)

The first two categories can be thought of as representations. The latter two make stronger claims: the model is now producing new answers or new behaviour and researchers may be tempted to treat those outputs as observations.

The key distinction: “AI organizes evidence we collected” is very different from “AI generates evidence we never collected.”

Sources: [3] Salminen et al.; [6] Wang & Siu; [1] Hämmäläinen et al.; [2] Xiang et al.

Epistemic Distance

The further we move right, the stronger the claim

Think in terms of epistemic distance from observed humans.



INCREASING EPISTEMIC DISTANCE FROM OBSERVED HUMANS →

The risk is not “synthetic-ness.” It is how much evidentiary work we ask the simulation to do. How far removed is this finding from something we actually observed a human being say or do?

Original synthesis. Conceptual framing informed by the empirical literature summarized later in the talk.

SPEAKER NOTES

Signature framework #1: Epistemic distance (2 minutes)

The left side is relatively modest: summarizing or representing evidence. The right side asks AI to make claims about human response or behaviour without direct observation.

“My concern isn’t really how synthetic the user is. It is how much epistemic work we are asking the simulation to do.”

Epistemology is the study of knowledge, how we know what we know, what counts as evidence, and when we can reasonably say something is true.

Epistemology asks: “How do we know this?”

So with synthetic users, the epistemological question is not just “Does this answer sound believable?” but **“What evidence do we have that it reflects real users?”**

The Seduction

And the appeal is obvious

FAST

seconds, not weeks

- No recruiting

- Instantaneous samples

CHEAP

no recruitment fees or incentives

- No specialty fees

- Any hour

SCALABLE

dozens → thousands

- Any number

- Any scenario

UXAgent (Lu et al., CHI EA '25) explicitly explores large-scale simulated users to evaluate study design before human-subject research.

SPEAKER NOTES

Why this is seductive (90 seconds)

Make the strongest possible case for the attraction before criticizing it. UX research has chronic pain points: access, time, incentives, repeated recruitment, and iteration cost. Synthetic users appear to dissolve all of them.

This is why a dismissive stance is not useful. If the method offers genuine leverage, we need to understand where it is valid.

Source: [7] Lu et al. UXAgent.

The Case For

Early evidence: surprisingly human-looking answers

Hämäläinen, Tavast & Kunnari — CHI 2023

Human response

"I usually play games to relax after work, especially when I don't have the energy for anything social."

LLM-generated response

"I tend to play in the evening because it helps me unwind and gives me something low-pressure to focus on."

The unsettling part:

It can sound completely plausible.

But "human-like text" is only one kind of similarity.

Conceptual illustration; cite Hämäläinen, Tavast & Kunnari (CHI '23), especially Fig. 5 and their human-likeness/diversity analyses.

SPEAKER NOTES

Hämäläinen et al. (2 minutes)

Explain that this CHI 2023 paper helped make synthetic HCI data a serious topic. Their experiments showed that language-model outputs could look impressively human-like on some surface-level measures.

The overarching takeaway from Hämäläinen et al. (2023) is a dual-use warning. LLMs can dramatically accelerate the design, prototyping and pilot testing phases of research by providing highly plausible synthetic data. However, relying on them to capture the genuine, highly diverse margins of human behavior will systematically erase the unique outliers that drive true user-centered innovation.

Transition: "Notice the standard we just used: plausible."

Source: [1] Hämäläinen et al. 2023. Suggested paper figure to cite in notes/handout: Fig. 5.

The Case For

Plausible doesn't mean representative

A model can imitate the style of a participant sample while compressing its diversity.



Model Convergence

fewer surprising,
idiosyncratic and
minority responses

A synthetic sample can be fluent while still being statistically and experientially "too smooth."

Hämäläinen et al. (CHI '23) found important diversity limitations despite human-like outputs; later literature repeatedly identifies synthetic homogeneity/distortion.

SPEAKER NOTES

Diversity problem (2 minutes)

Explain model convergence: **language models optimize likely continuations**, so a synthetic "sample" can sound varied while still over-producing typical or semantically central responses. By definition this gravitates towards a central tendency.

This matters because research is often valuable precisely because of the unexpected, awkward, minority, and contradictory observations—not because of the average answer.

Sources: [1] Hämäläinen et al.; [12] Kuric et al. systematic review. Suggested paper item to cite: Hämäläinen et al.'s diversity analyses/table rather than copying a figure.

From Simulation to Interaction

SimUser: What about usability testing

When Synthetic Users Start Clicking

SimUser generates heuristic usability feedback by simulating different users interacting with a mobile interface.



Redrawn conceptually from Xiang et al. (CHI '24), Fig. 1 / system overview.

SPEAKER NOTES

SimUser setup (2 minutes)

Explain why this study raises the stakes. Rather than generating interview answers, the system attempts to simulate user–application interaction and generate usability feedback.

Source: [2] Xiang et al. 2024. Based on Fig. 1/system overview.

The capability emerged in stages, but **2022–2024 is the key transition period.**

A useful timeline:

- **2022:** Adept demonstrated **ACT-1**, a model that could use a web browser by observing the UI and taking actions such as clicking, typing, and scrolling. This was an early, high-profile demonstration that language models could operate existing software interfaces rather than only call APIs.
- **2023:** Research systems such as **WebArena** began systematically benchmarking LLM agents on realistic,..(Continued - next slide)

From Simulation to Interaction

SimUser: What about usability testing

When Synthetic Users Start Clicking

SimUser generates heuristic usability feedback by simulating different users interacting with a mobile interface.



Redrawn conceptually from Xiang et al. (CHI '24), Fig. 1 / system overview.

SPEAKER NOTES

- ..multi-step web tasks across functioning websites. Performance was still poor—its best GPT-4-based agent completed only about **14%** of tasks end-to-end—but the research direction was clearly established.
- **2024:** This became much more concrete. On **October 22, 2024**, Anthropic released Claude 3.5 Sonnet's **computer use** capability in public beta, explicitly allowing the model to look at screens, move the cursor, click buttons, and type into ordinary applications. Anthropic described it as the first frontier model to offer general computer use in public beta.
- **2025:** OpenAI launched **Operator / Computer-Using Agent** on January 23, 2025. It operated through screenshots plus mouse and keyboard actions and was designed to use ordinary GUIs without requiring application-specific APIs. (Continued)

From Simulation to Interaction

SimUser: What about usability testing

When Synthetic Users Start Clicking

SimUser generates heuristic usability feedback by simulating different users interacting with a mobile interface.



Redrawn conceptually from Xiang et al. (CHI '24), Fig. 1 / system overview.

SPEAKER NOTES

- **2025–2026:** Computer-use capabilities moved from demos and research previews toward more general-purpose agent products and developer tools, with systems increasingly able to execute real tasks across websites and desktop software.

GUI-using agents did not suddenly appear after Xiang et al. The capability was already emerging. But 2024 was the inflection point when computer use began moving from research prototypes into general-purpose frontier models, and 2025 made it a user-facing capability.

It found real problems, but performance varied dramatically

SimUser evaluation: 48 human users + 21 designers on a simple smartwatch interface.

35.7%

LOW END



100%

HIGH END

Similarity of generated usability feedback to human feedback, depending on user group and usability category

PROMISING

NOT EQUIVALENT

Xiang et al. (CHI '24): "similarity varying from 35.7% to 100% according to the user groups and usability category."

SPEAKER NOTES

SimUser results (2 minutes)

A synthetic tool that identifies a substantial fraction of human-identified usability feedback is useful. The issue is variability: 35.7% and 100% are very different levels of fidelity.

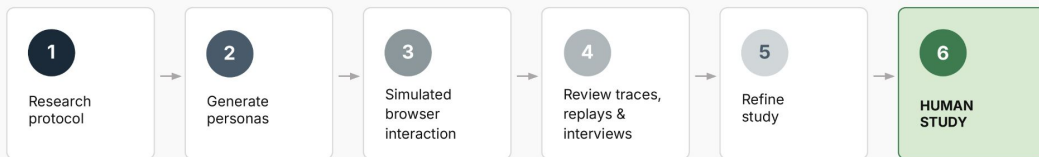
The paper itself frames SimUser as generating heuristic usability feedback, which is a more cautious and defensible role than replacing participants.

Source: [2] Xiang et al. 2024. The 35.7–100% range is reported in the paper abstract/evaluation summary.

Grounded Agents

Perhaps the better question is: what should we automate?

UXAgent — Lu et al., CHI 2025



Use simulation to find flaws in the study before recruiting real participants.

UXAgent is explicitly positioned as support for evaluating and iterating a usability study design before the real human-subject study.

SPEAKER NOTES

UXAgent (about 2 minutes)

Using Agents for pilot testing could be the safe middle ground.

This is one of the most defensible applications. UXAgent uses generated personas and web agents to expose possible problems with the study design itself: unclear tasks, broken flows, missing probes, etc.

AI isn't replacing the participant here. It is helping the researcher avoid wasting a participant.

Source: [7] Lu et al. 2025. Cite Fig. 1/system design if desired.

Grounded Agents

Grounding changes the equation

51

Knowledge workers

3,060

human concept-test responses

4

novel AI concepts

DISTRIBUTION-CALIBRATED

but

IDENTITY-IMPRECISE

Population-level distributions could look similar while the “digital twin” often failed to reproduce the specific person it represented.

Wang & Siu (CHI '26), Fig. 1 and study overview; abstract conclusion: “distribution-calibrated but identity-imprecise.”

SPEAKER NOTES

Grounded agents (2–3 minutes)

This study from **Adobe Research** is important because the agents were not just demographic prompts. Researchers first interviewed the actual participants about their workflows, then created personalized agents, then compared the agents’ concept evaluations with the same humans.

90 minute interviews with 51 knowledge workers, then presented them with 4 novel AI concepts. They were asked to rate the concepts for perceived usefulness, ease of use and Net Promoter Score, plus written open ended feedback. The constructed a unique LLM clone for each participant.

This is similar to the concept of a digital twin. Which is an intriguing concept, but may have ethical implications.

The headline finding is wonderfully nuanced: “distribution-calibrated but identity-imprecise.” At population level, the simulation can be directionally useful. At individual level, digital Maria is not reliably Maria.



Grounded Agents

Grounding changes the equation

51

Knowledge workers

3,060

human concept-test responses

4

novel AI concepts

DISTRIBUTION-CALIBRATED

but

IDENTITY-IMPRECISE

Population-level distributions could look similar while the “digital twin” often failed to reproduce the specific person it represented.

Wang & Siu (CHI '26), Fig. 1 and study overview; abstract conclusion: “distribution-calibrated but identity-imprecise.”

SPEAKER NOTES

..how closely does the 1-7 likert score match the human counterpart. The best clones only managed to match 67% of the human baseline. When you zoom out and look at the aggregated data, the shape of the AI data curve mirrored the shape of the human generated curve. Think of weather... meteorologist can tell you it rains 30% of the time in September in Ottawa. This can be done very accurately. Will they be able to tell you if it will rain at 2:00pm on Saturday while you're apple picking... no.

Source: [6] Wang & Siu 2026. Cite Fig. 1 and the abstract/conclusion.

\$12.50 per human user/30 minutes vs. \$4 minutes \$1.27 eliminates the logistics of recruiting and scheduling.

The “no free participant” paradox

The strongest simulations often depend on human research upstream.



The simulation becomes strongest when human evidence already exists.

Original synthesis based on Wang & Siu (CHI '26) and Park et al. (2024), both of which ground agents in real qualitative interview data.

SPEAKER NOTES

Signature framework #2 (2 minutes)

Introduce this as your own phrase: “the no free participant paradox.”

The better the simulation, the more often we discover that its performance derives from real participants upstream. Wang & Siu rely on workflow interviews; Park et al. build agents from long interviews with 1,052 people.

This changes the business case. Maybe the right role is not replacing research, but extending the shelf-life and reach of research we have already paid to conduct.

Sources: [6] Wang & Siu; [8] Park et al.

Where the case is strongest

Use synthetic users to create possibilities, not to manufacture findings.

Generate possibilities — not findings.

✓ Generate hypotheses / questions

GOOD FIT

✓ Stress-test a protocol

GOOD FIT

✓ Heuristic design critique

GOOD FIT

✓ Interrogate existing research

GOOD FIT

Synthesis consistent with the augmentative framing in UXAgent, SimUser, Wang & Slu, and the 2026 systematic review.

SPEAKER NOTES

Interim synthesis (90 seconds)

This is the most positive slide in the talk. All of these uses reduce researcher effort without requiring you to pretend that model output is an observation of a person.

Stress: some of these applications may become extremely valuable in practice. This is not an anti-synthetic-user argument; it is a role-definition argument.

Sources: [2], [6], [7], [12].

You are now the research participants

Audience experiment: first-click expectation

TASK

You recently moved. Where would you click first to change the mailing address associated with an existing parking permit?

Don't optimize. Click where you would naturally go first.

CITY SERVICES

My account

- My permits
- Profile & contact information
- Account settings
- Parking & vehicles
- Help

SPEAKER NOTES

Audience experiment setup (90 seconds)

Tell the room that there is intentionally no obviously correct answer. The dependent variable is first-click expectation, not task completion.

If using Mentimeter/Slido, have the poll ready and show the QR code before advancing. If not, ask for hands for the five choices.

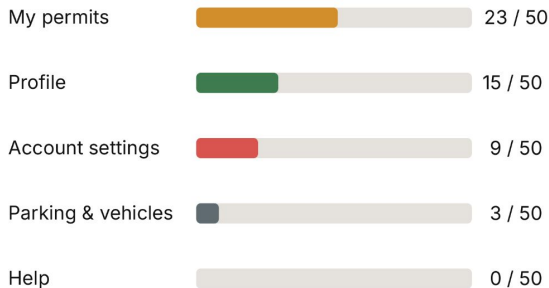
This fictional interface avoids contaminating the demo with known brand conventions or an existing government site.

INSERT SLIDO

Here are 50 people who don't exist

When human and synthetic distributions disagree, who is wrong?

SYNTHETIC PANEL (N=50)



**More synthetic
samples
≠
more human
evidence.**

Company Confidential ©OFFRINT Inc. All Rights Reserved.

SPEAKER NOTES

Reveal and discuss (about 3 minutes)

IMPORTANT: Replace the right-side illustrative panel before the talk. Pre-run exactly the same task/image with your chosen model and record the model/version, date, system prompt, task prompt, sampling settings, and number of independent runs. Keep this slide's tiny "demo data" label until you replace it.

Ask: "Suppose the synthetic sample has $n=5,000$ and this room has $n=40$. Does that make the synthetic result better evidence?"

The point is epistemic, not statistical: larger generated N does not create more independent observations of the human population.

Section Title

Plausibility ≠ validity

PLAUSIBILITY

≠

VALIDITY

“Does this sound like something a human might say?”

“Does this tell us what this population actually thinks or does?”

LLMs can pass the first test while failing the second.

Kapania et al. (CHI '25) and Kuric et al. (2026 SLR) both highlight the danger of believable outputs that lack the properties of human evidence.

SPEAKER NOTES

Central methodological claim (about 2 minutes)

Slow down here. This is the intellectual centre of the talk.

A model can be judged on face validity: “this sounds believable.” UX research requires something stronger: construct, external, and conclusion validity about actual people.

Kapania et al. show why qualitative researchers may initially find generated narratives plausible but then identify missing context, agency, and situated experience. The systematic review calls the problem “misleading believability.”

Sources: [5] Kapania et al.; [12] Kuric et al.

2026 Preprint

What would GPT click?

Kuric, Demcak & Krajcovic — 12 real first-click studies, 3,431 human participants

53%

of tasks: significantly different
human vs synthetic distributions

71%

same most-popular first click

.16

mean rank-order agreement

DIVERSITY OF FIRST CLICKS



Company Confidential ©DIFFRINT Inc. All Rights Reserved.

Kuric et al. 2026, Table 3 and moderation analyses. Preprint: arXiv:2605.18302.

SPEAKER NOTES

First-click evidence (2–3 minutes)

Label the work clearly as a preprint. The strength is ecological validity: 12 real UXtweak first-click studies with 3,431 human participants, using actual visual stimuli.

The headline “53% significantly different” matters, but the deeper point is distributional shape. The same winner in 71% of tasks can sound good, yet mean rank-order agreement is only .16. Humans also showed more diverse clicks: normalized entropy .61 vs .43 and about 9 unique selections vs 5.38 for baseline synthetic participants.

Source: [10] Kuric et al. 2026, especially Table 3.



Section Title

Maybe we just need better prompting?

The familiar escape hatch: model, reasoning, temperature, top-p, personas...

Better Model

Top-p

Chain-of-thought

Detailed personas

Temperature

No Systemic Rescue.

SINGLE-PERSONA CONDITION

96% of tasks still had different click distributions.

“Prompt it better” is not a substitute for empirical validation.

Company Confidential ©DIFFRINT Inc. All Rights Reserved.

Kuric et al. 2026, Table 3 and moderation analyses. Preprint: arXiv:2605.18302.

SPEAKER NOTES

Prompting interventions (about 2 minutes)

Walk through the obvious proposed fixes: better models, chain-of-thought, sampling changes, detailed personas. In this first-click dataset, those changes did not turn synthetic participants into reliable human behavioural proxies.

The single-persona condition is especially useful pedagogically: it produced different distributions in 96% of tasks and collapsed click diversity.

Do not generalize this to “prompting never helps.” The correct claim is narrower: in this evaluation, prompting did not rescue human fidelity.

Source: [10] Kuric et al. 2026, Table 3.



2026 Preprint

What about design preferences

29 real studies · 2,073 people · 78 design tasks

44%

tasks with different human vs
synthetic distributions

53%

same first-choice design

.93 vs .86

synthetic vs human normalized
entropy

Synthetic panels were often more evenly indecisive than humans — even when humans had a clearer favourite.

Different failure mode, same problem: distribution ≠ population.

Company Confidential ©DIFFRNT Inc. All Rights Reserved.

Kuric et al. 2026 preference study, qualitative justification analysis; Kapania et al. (CHI '25) on contextual depth and surrogate narratives.

SPEAKER NOTES

Preference testing (2–3 minutes)

Again, label as preprint. This companion study uses 29 real preference tests from UX practice. Across 78 tasks, 44% had significantly different choice distributions and only 53% shared the same first choice.

A useful contrast with the click study: synthetic click behaviour was often too concentrated, whereas synthetic design preference was often too evenly distributed (entropy .93 synthetic vs .86 human).

The lesson is not simply “synthetic users lack diversity.” They can distort diversity in different directions depending on task.

Source: [11] Kuric et al. 2026, Results §5.1.



Section Title

But the explanations sound so thoughtful...

Believable rationales can increase confidence without increasing fidelity.

GENERIC

"clear / intuitive / clean"

OVER-ELABORATE

longer than the human answer needs to be

OVER-POSITIVE

praise where humans are mixed or critical

CONFIDENTLY IRRELEVANT

a polished reason for the wrong thing

The Plausibility Trap

Company Confidential ©DIFFRINT Inc. All Rights Reserved.

Kuric et al. 2026 preference study, qualitative justification analysis; Kapania et al. (CHI '25) on contextual depth and surrogate narratives.

SPEAKER NOTES

Qualitative reasoning (2 minutes)

The paradox: explanations can make weak simulations more dangerous because explanation increases trust. Kuric et al. find recurring synthetic patterns such as generic design-property talk, over-elaboration, praise, and contextually nonsensical reasoning.

Kapania et al. reaches a related conclusion in qualitative research: generated stories can be compelling while missing situated context, agency, and lived complexity.

Call this "the plausibility trap."

Sources: [11] Kuric et al.; [5] Kapania et al.

Section Title

Nobody was frustrated

A generated report of experience is not an observed experience.

HUMAN PARTICIPANT

- Perception
- Attention
- Memory
- Motor action
- Past experience
- Fatigue
- Emotion
- Social context
- Consequences
- Lived experience

SYNTHETIC PARTICIPANT

- Prompt
- Context window
- Training data
- Model architecture
- Sampling
- Generated response

**“I felt frustrated” is
not an observation
of frustration.**

Company Confidential ©DIFFRINT Inc. All Rights Reserved.

Methodological synthesis informed by Kapania et al. (CHI '25) and broader simulation-validity literature.

SPEAKER NOTES

Title payoff (about 2 minutes)

Be careful not to turn this into a debate about machine consciousness. Say explicitly: “I’m making a research-method claim, not a metaphysical claim.”

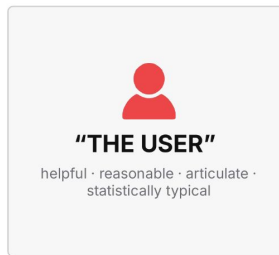
No human participant had the reported experience. Therefore a generated sentence saying “I was frustrated” is not evidence that a participant experienced frustration. It may be a useful hypothesis about where frustration could occur. Those are different epistemic statuses.

Source: [5] Kapania et al. is especially useful for grounding the distinction between generated narrative and situated human account.

Section Title

The “average user” comes back

We spent decades trying to stop designing for a mythical average user.



Synthetic participants risk turning population diversity into a polished central tendency.

Company Confidential ©DIFFRINT Inc. All Rights Reserved.

Salminen et al. (CHI '24) document compositional biases in 450 AI-generated personas; Agnew et al. (CHI '24) critique "artificial inclusion."

SPEAKER NOTES

Bias and inclusion (2 minutes)

This is where the ethical and methodological problems meet. Persona generation can introduce compositional biases in nationality, age, occupation, gender, etc. Salminen et al. tested 450 generated personas and found systematic composition problems.

Agnew et al.'s "Illusion of Artificial Inclusion" is useful as a normative counterpoint: simulating a marginalized perspective is not the same as including people who hold that perspective in the research process.

Avoid asserting that every synthetic-user system collapses to an average user. Frame it as a recurring risk.

Sources: [3] Salminen et al.; [4] Agnew et al.

So, where does the evidence stand?

Kuric, Demcak & Krajcovic synthesize 182 studies

1. Cognitive misalignment

2. Systemic distortion

3. Misleading believability

4. Overfitting/contamination

USEFUL ≠ EQUIVALENT

Company Confidential ©DIFFRINT Inc. All Rights Reserved.

Kuric, Demcak & Krajcovic (Research Square preprint, Mar. 2026): review of 182 studies; proposes a "heuristic-like" framing.

SPEAKER NOTES

Systematic review (2 minutes)

Label it as a preprint and summarize rather than oversell. The review synthesizes 182 studies and organizes recurring issues into four categories: cognitive misalignments, distortions, misleading believability, and overfitting/contamination.

Its framing is useful because it does not say "synthetic participants are useless." It proposes treating them as heuristic-like tools, with modest fidelity gains from model/prompt/persona interventions across the literature.

A very recent July 2026 cross-domain survey benchmark by Chen et al. provides another cautionary example: demographic prompting can over-determine segment differences and misdirect targeting decisions.

Sources: [12] Kuric et al.; [13] Chen et al.



Section Title

The synthetic user validity ladder

A practical way to match synthetic methods to evidence claims.

Strong use

Generate hypotheses · Stress-test protocols · Explore edge cases · Heuristic critique

Conditional

Converse with existing research · Extend grounded evidence · Early concept screening after validation

Weak evidence

Predict attitudes · Predict usability issues · Predict individual behaviour

Do not treat it as a human evidence

Estimate prevalence without calibration · Replace validation · Claim lived experience · Simulate inclusion

Company Confidential ©OFFPRINT Inc. All Rights Reserved.

SPEAKER NOTES

Signature framework #3 (2–3 minutes)

This is your practical contribution. The ladder is not a universal standard—it is a decision aid built from the evidence in the talk.

The rule underneath it: “Use synthetic users to create things to test—not to manufacture confirmation that those things are true.”

Explain why “conditional” includes grounded concept screening: Wang & Siu show a narrow case where distributional calibration can support directional early-stage screening, but authentic interviews remain essential for individual workflows, trust, and barriers.

Sources: [1], [2], [5], [6], [7], [10]–[13].



Section Title

Six rules for synthetic-user research

A defensible operating model for UX teams

1. Name the epistemic role

Hypothesis generator? Heuristic evaluator?
Predictor? Interface to evidence?

2. Ground where possible

Use real interviews, logs or research repositories instead
of demographic role-play alone.

Synthetic users are
research instruments.
They are not research
participants.

**The danger is that their
answers become
convincing enough that we
stop caring about the
distinction.**

3. Validate in-domain

Fidelity is task- and context-specific. Do not inherit
validity from another study.

4. Match evidence to consequence

The higher the stakes, the less room for unvalidated
simulated evidence.

5. Investigate disagreement

Human-synthetic disagreement is a finding to
study, not noise to average away.

6. Disclose provenance

Model, version, date, prompts, grounding, sampling and
human comparator.

Company Confidential ©DIFFRINT Inc. All Rights Reserved.

SPEAKER NOTES

Conclusion (2–3 minutes)

Return to Slide 3: “Was my study of 300
nonexistent customers user research?”

Your answer: “I think the answer is no. But it might
have been an extraordinarily useful research
instrument.”

Then deliver the closing distinction: “Synthetic
users are research instruments. They are not
research participants.”

Final line: “The danger isn’t that synthetic users
will become indistinguishable from real users. The
danger is that their answers will become
convincing enough that we stop caring about the
distinction.”

Stop. Do not add another summary slide. Leave
the reference slides up for Q&A.



References

Recent simulation studies and 2026 preprints used in the talk

- [1] Perttu Hämmäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). ACM. <https://doi.org/10.1145/3544548.3580688>
- [2] Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). ACM. <https://doi.org/10.1145/3613904.3642481>
- [3] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J. Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). ACM. <https://doi.org/10.1145/3613904.3642036>
- [4] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). ACM. <https://doi.org/10.1145/3613904.3642703>
- [5] Shivani Kapania, William Agnew, Motahhare Eslami, Hoda Heidari, and Sarah E. Fox. 2025. Simulacrum of Stories: Examining Large Language Models as Qualitative Research Participants. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). ACM. <https://doi.org/10.1145/3706598.3713220>
- [6] Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Zheshen Jessie Wang, Yang Li, Jiri Gesi, Qi He, Toby Jia-Jun Li, and Dakuo Wang. 2025. UXAgent: An LLM Agent-Based Usability Testing Framework for Web Design. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25). ACM. <https://doi.org/10.1145/3706599.3719729>

References

Recent simulation studies and 2026 preprints used in the talk

- [7] Zichao Wang and Alexa Siu. 2026. Interview-Informed Generative Agents for Product Discovery: A Validation Study. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26). ACM. <https://doi.org/10.1145/3772318.3791918>
- [8] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109. <https://doi.org/10.48550/arXiv.2411.10109>
- [9] Huyen N. Nguyen, Astrid van den Brandt, and Nils Gehlenborg. 2026. Sycamore: Characterizing Synthetic Personas for Evaluating Genomics Visualization Retrieval. arXiv:2605.08630.
- [10] Eduard Kuric, Peter Demcak, and Matus Krajcovic. 2026. What Would GPT Click: Practical Effects of Human-AI Behavioral Misalignment and the Cost of Synthetic Participants in User Experience. arXiv:2605.18302.
- [11] Eduard Kuric, Peter Demcak, and Matus Krajcovic. 2026. Distorted Perspectives of LLM-Simulated Preferences: Can AI Mislead Design? arXiv:2605.18311.
- [12] Eduard Kuric, Peter Demcak, and Matus Krajcovic. 2026. Synthetic Participants Generated by Large Language Models: A Systematic Literature Review. Research Square preprint. <https://doi.org/10.21203/rs.3.rs-9057643/v1>
- [13] Zihan Chen, Di Zhu, and Lei Nico Zheng. 2026. When Synthetic Users Fail: A Cross-Domain Benchmark of LLM-Simulated Human Survey Responses. arXiv:2607.26348. Manuscript under peer review.

Note: slides explicitly label 2026 preprints as preprints. Model capabilities and this evidence base are moving rapidly; re-check status before presenting.

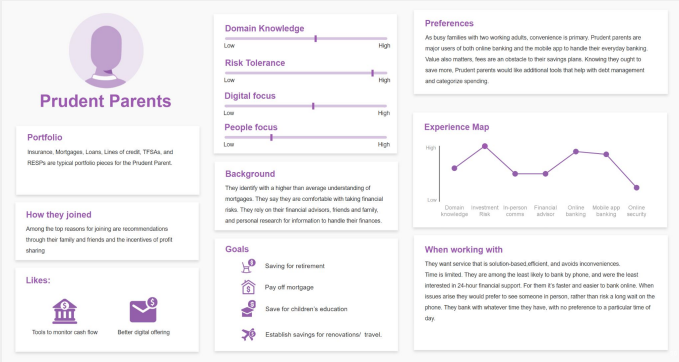
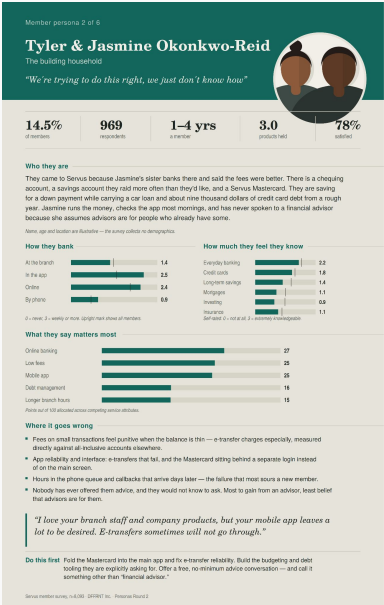
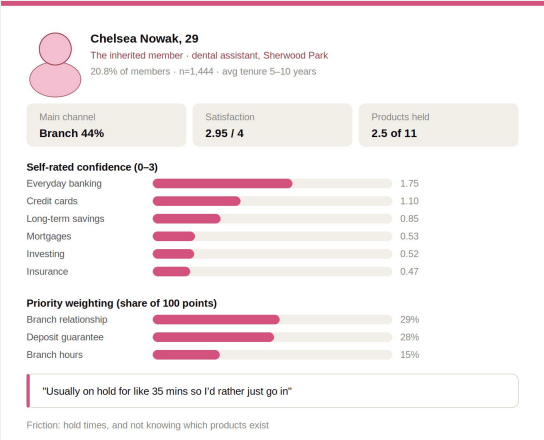
Appendix 1: Synthetic Personas

Appendix 1 - Synthetic Personas

Experiment 1: Synthetic Personas

Comparing Claude’s persona creation to human UX researcher outputs

We took a survey dataset and asked Claude to create personas from it - once without behavioural survey questions, and once with.

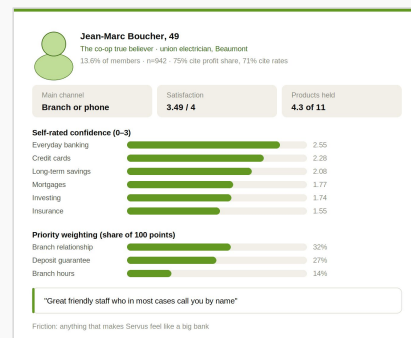
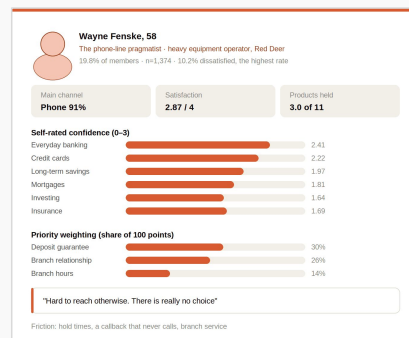
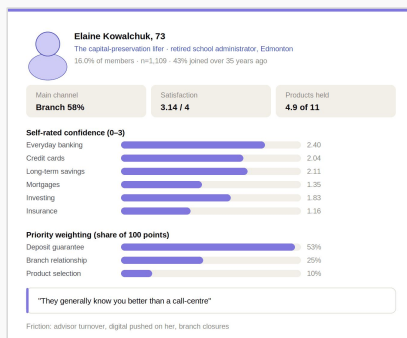
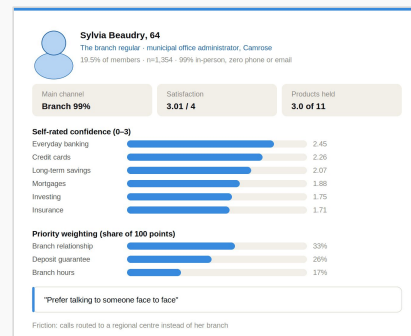
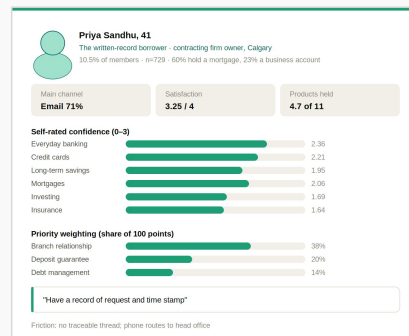
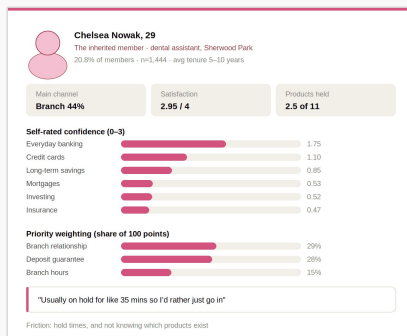


Appendix 1 - Synthetic Personas

Claude attempt 1: No behavioural questions and quotes included

6 personas; lack of clear descriptive distinction between each; confidence almost identical for each; design of bar charts weak (33% taking up half the bar!).

Overall a poor output - but fast.
This took 15 minutes to create after simply uploading survey results



Appendix 1 - Synthetic Personas

Claude attempt 2:
behavioural questions
and quotes included

6 personas; much more developed;
proper descriptions and quotes;
clear goals and motivations; visuals
more pleasing (though still in need
of human editing).

Overall output: Moderately good,
and *fast*: This again only took about
15 minutes to produce

Member persona 1 of 6

Marlene Kowalchuk

The long-tenured, advisor-led saver

"I've been here since it was Capital City"

16.7%

of members

1,110

members

35+ yrs

a member

5.1

products held

90%

satisfied

Who she is

Marlene opened her first account at sixteen, when the branch still had a different name over the door, and retired two years ago after thirty-one years with the school division. Her money sits in RRSPs, a GIC ladder she renews every spring, and a modest mutual fund portfolio she trusts her advisor to steer. She reads her statements on paper. The app is on her phone because her daughter installed it, and she has opened it perhaps four times.

Name, age and location are fictitious — the survey collects no demographics.

How they bank

At the branch 17

In the app 1.4

Online 2.8

By phone 0.8

0 = never, 3 = weekly or more (longer rank shows of members)

How much they feel they know

Everyday banking 2.4

Credit cards 2.1

Long-term savings 2.2

Mortgages 1.3

Insurance 1.2

Insurance (offered 0 = not at all, 3 = exclusively via mortgage)

What they say matters most

100% deposit guarantee 40

Knowledge and advice 31

Branch relationship 31

Staff easy to work with 31

Online banking 31

Points out of 100 allocated across competing service attributes.

Where it goes wrong

Advisor churn breaks the relationship she is paying for — after a round of staff changes she does not know who her advisor is.

Calls to her own branch route to a contact centre, which she experiences as being cut off from the people she knows by name.

Being told to book an appointment to ask a question reads as a downgrade in service.

Tellers offering investment advice she then has to pay to unfund.

Advisor appointments and price giveaways read as her profile share being spent elsewhere.

"Now that ALL the staff has changed in the past few weeks at my branch, who is my Financial Advisor?"

Do this first: Named-advisor continuity with a warm handover when staff change. Direct-to-branch phone routing. Lead with deposit guarantee and member ownership — and be able to explain how profit share is calculated.

Source: member survey, n=6,000 | ©FFRNT Inc. | Persona Board 2

Member persona 2 of 6

Tyler & Jasmine Okonkwo-Reid

The building household

"We're trying to do this right, we just don't know how"

14.5%

of members

969

members

1-4 yrs

a member

3.0

products held

78%

satisfied

Who they are

They came to Servus because Jasmine's sister banks there and said the fees were better. There is a chequing account, a savings account they raid more often than they'd like, and a Servus Mastercard. They are saving for a down payment while carrying a car loan and about nine thousand dollars of credit card debt from a rough year. Jasmine runs the money, checks the app most mornings, and has never spoken to a financial advisor because she assumes advisors are for people who already have some.

Name, age and location are fictitious — the survey collects no demographics.

How they bank

At the branch 1.4

In the app 2.8

Online 2.4

By phone 0.5

0 = never, 3 = weekly or more (longer rank shows of members)

How much they feel they know

Everyday banking 2.2

Credit cards 1.8

Long-term savings 1.4

Mortgages 1.1

Insurance 1.1

Insurance (offered 0 = not at all, 3 = exclusively via mortgage)

What they say matters most

Online banking 27

Low fees 25

Mobile app 25

Debt management 16

Longer branch hours 15

Points out of 100 allocated across competing service attributes.

Where it goes wrong

Fees on small transactions feel punitive when the balance is thin — e-transfer charges especially, measured directly against all inclusive accounts elsewhere.

App reliability and interface: e-transfers that fail, and the Mastercard sitting behind a separate login instead of on the main screen.

Hours in the phone queue and callbacks that arrive days later — the future that most secure a new member.

Nobody has ever offered them advice, and they would not know to ask. Most to gain from an advisor, least belief that advisors are for them.

"I love your branch staff and company products, but your mobile app leaves a lot to be desired. E-transfers sometimes will not go through."

Do this first: Fold the Mastercard into the main app and fix e-transfer reliability. Build the budgeting and debt-busting they are explicitly asking for. Offer a live, no-minimum advice conversation — and call it something other than "financial advisor."

Source: member survey, n=6,000 | ©FFRNT Inc. | Persona Board 2

Member persona 3 of 6

Dwayne Fournier

The mortgage-servicing homeowner

"Ten years in and still fifteen to go"

17.1%

of members

1,140

members

11-20 yrs

a member

5.5

products held

86%

satisfied

Who he is

Dwayne has a mortgage, a line of credit against the truck, RRSPs he tops up in February, an RESP for the older one, and a small business account for the wedding he does on weekends. He is the household's financial operator: he knows the rate, the renewal date, and what his amortisation looks like if he throws an extra two hundred a month at it. He banks on the app at nine in the evening, goes in for anything involving a rate or a signature, and would rather email his account manager than sit in a phone queue.

Name, age and location are fictitious — the survey collects no demographics.

How they bank

At the branch 1.8

In the app 2.2

Online 2.3

By phone 0.8

0 = never, 3 = weekly or more (longer rank shows of members)

How much they feel they know

Everyday banking 2.4

Credit cards 2.2

Long-term savings 1.9

Mortgages 2.2

Insurance 1.4

Insurance (offered 0 = not at all, 3 = exclusively via mortgage)

What they say matters most

Branch relationship 31

Online banking 31

Security 28

Low fees 25

Mobile app 19

Points out of 100 allocated across competing service attributes.

Where it goes wrong

Credit decisions that stall or get declined, sending him to another lender — the fastest route to losing a 5.5-product relationship.

His account manager is unreachable; weeks pass without a callback on a mortgage or line-of-credit file.

The Mastercard portal is a separate, worse system bolted onto the side of his banking.

Branch and ATM consultations to smaller cities: "missed downtown" means a twenty-minute drive for cash.

Advisor availability does not fit a working week — no evenings, no Saturdays.

"Not impressed with financial support. Whenever approached for business or personal loans, never able to come through. Always have to use other lenders."

Do this first: Named account-manager reachability with a callback SLA, and transparent lending decisions with a reason and a path. Mortgage left-serves in the app: renewal dates, prepayment modelling, amortisation what-ifs. Evening and Saturday appointments.

Source: member survey, n=6,000 | ©FFRNT Inc. | Persona Board 2

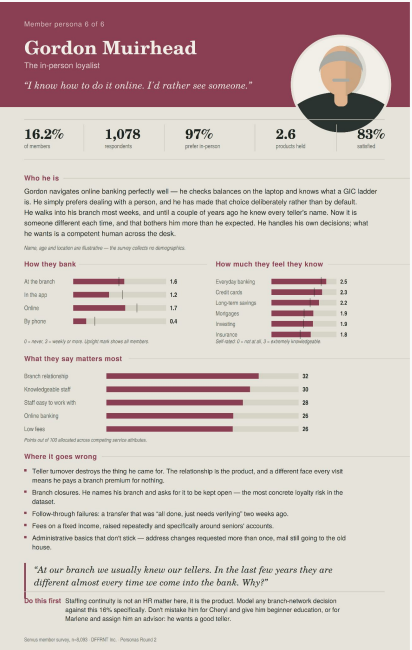
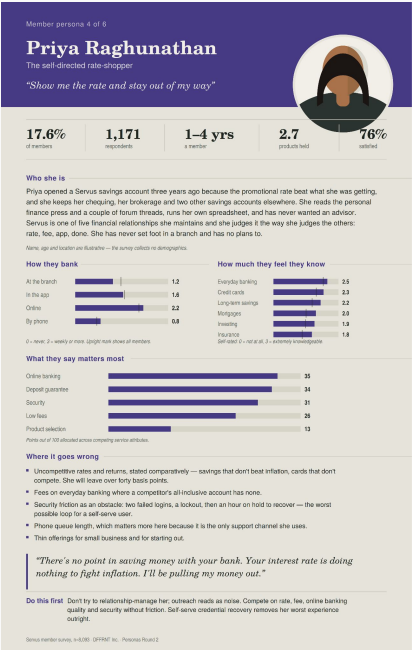
Company Confidential ©FFRNT Inc. All Rights Reserved.

Appendix 1 - Synthetic Personas

Claude attempt 2:
behavioural questions
and quotes included

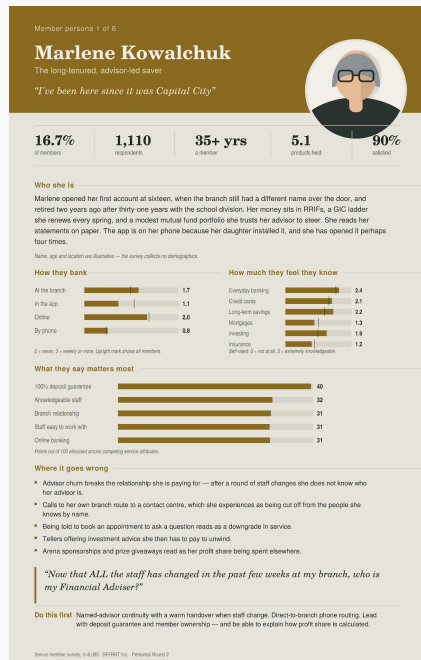
6 personas; much more developed;
proper descriptions and quotes;
clear goals and motivations; visuals
more pleasing (though still in need
of human editing).

Overall output: Moderately good,
and *fast*: This again only took about
15 minutes to produce



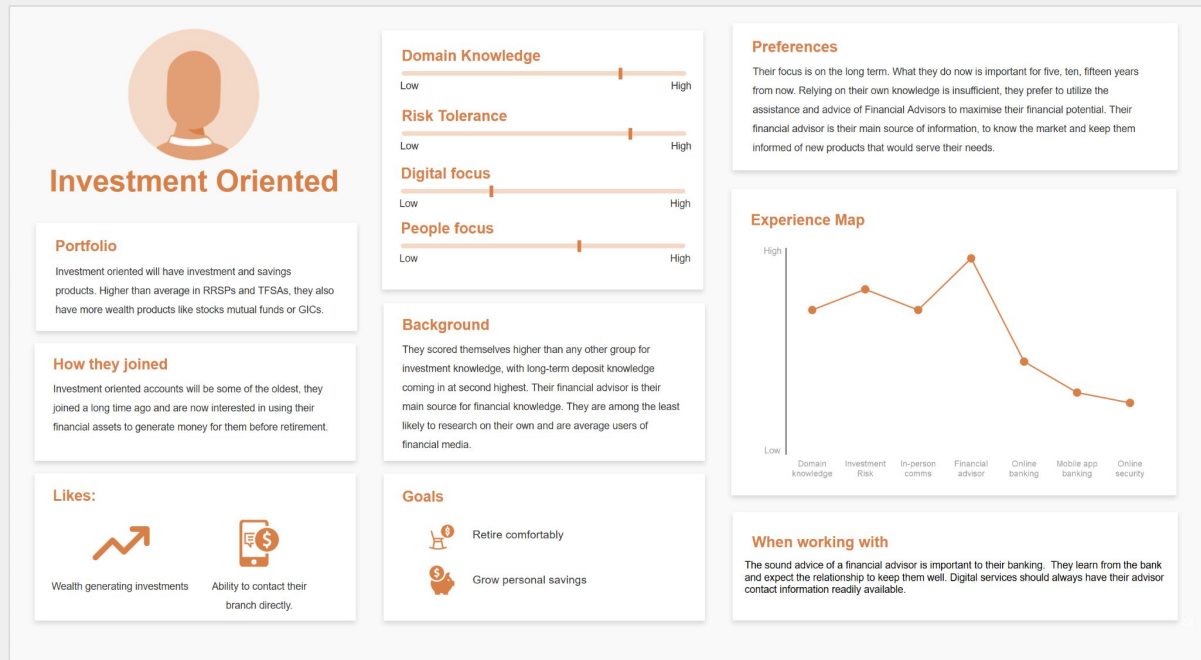
Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's



Claude-generated

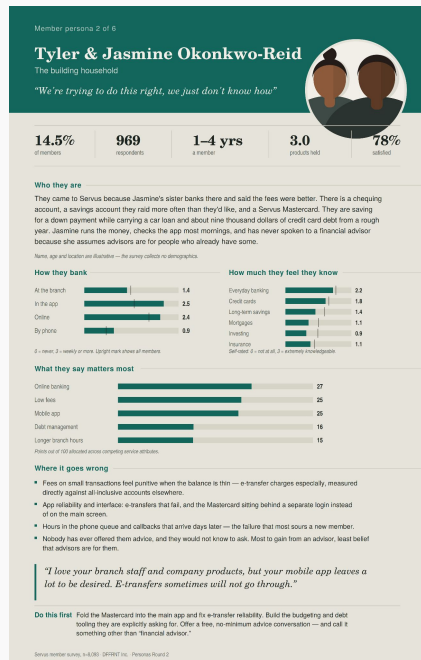
Near-identical: oldest accounts, advisor as primary information source, least likely to self-research, highest self-rated investing knowledge, RRSP/TFSA/wealth/GIC, and both flag "wants direct contact with their branch" as the design implication



Human-generated

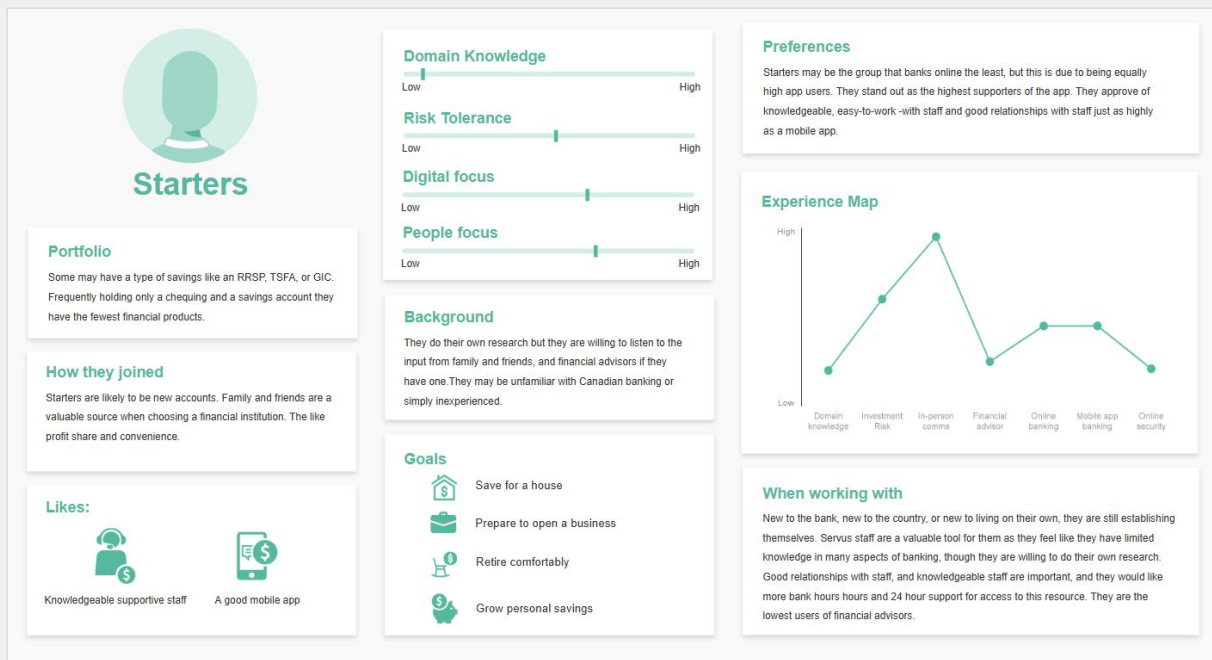
Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's



Claude-generated

Near-identical, including the odd specifics: newest accounts, highest app support, fewest products, lowest advisor use, referral by family and friends, longer branch hours and 24-hour support both peaking, and "prepare to open a business" as a real goal

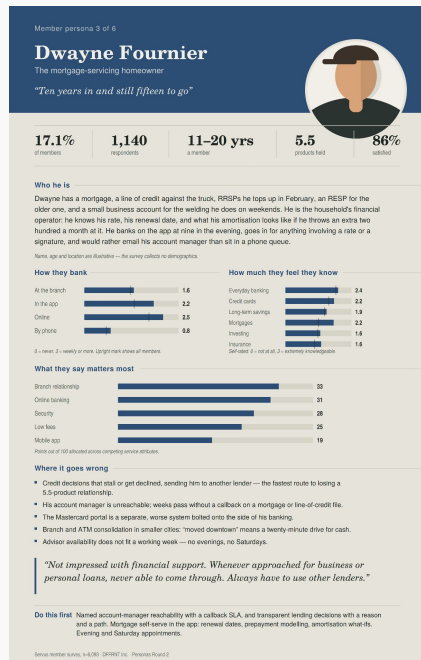


Human-generated

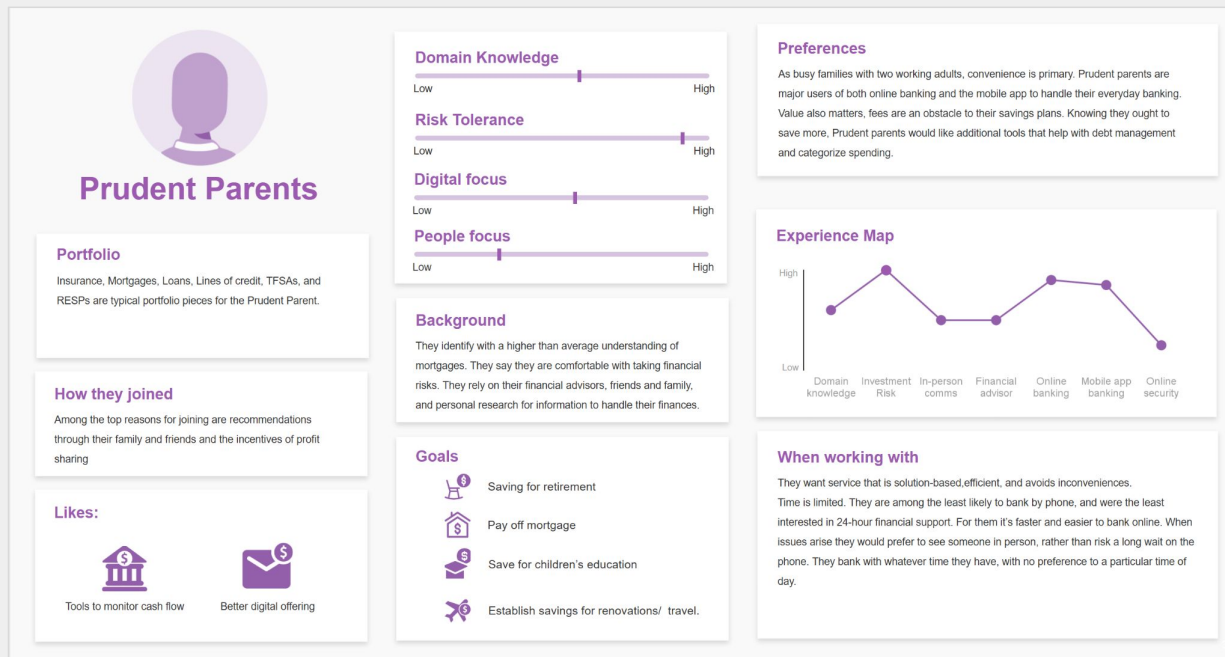
Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's

Strong similarity: above-average mortgage knowledge, fluent online and app, most risk-tolerant of the set, RESP and education goals, and explicitly asking for debt-management and spend-categorisation tools



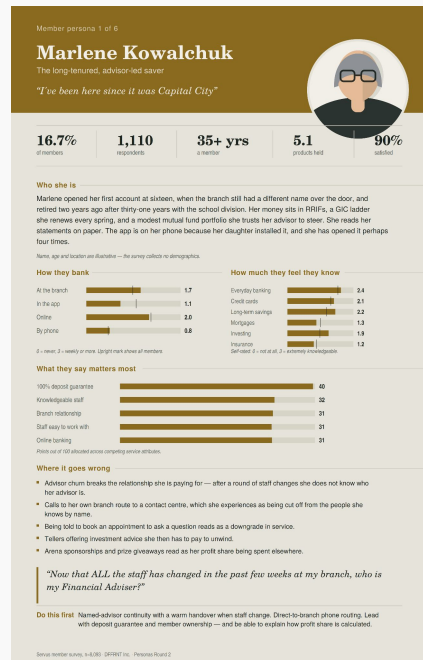
Claude-generated



Human-generated

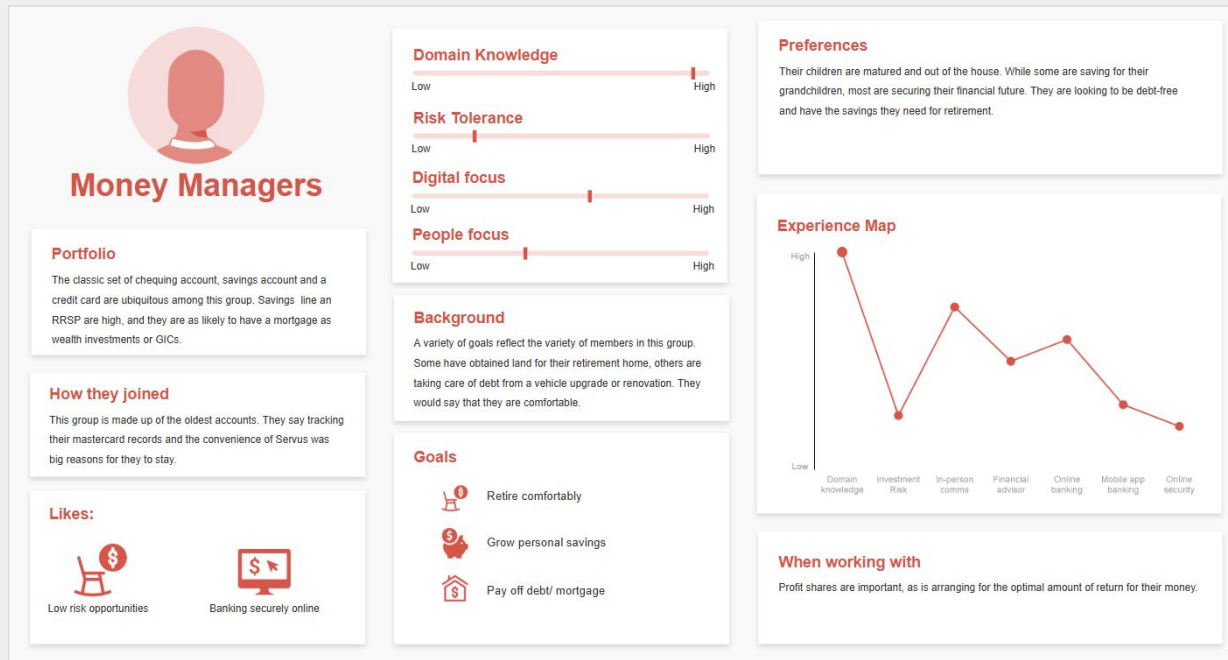
Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's



Claude-generated

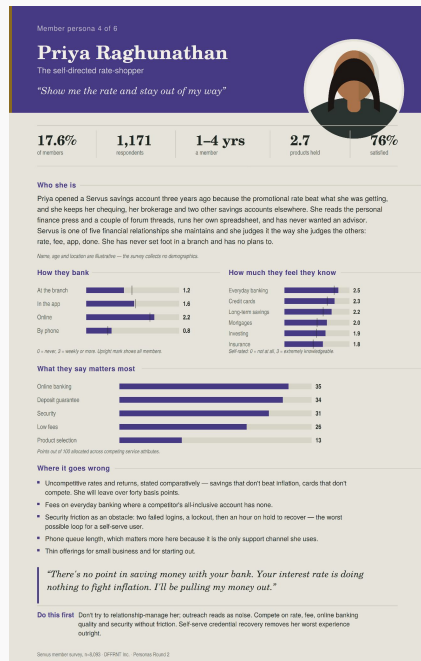
Partial fit: Retirement-securing and oldest accounts is Marlene; the Mastercard-record joining motivation is Dwayne's peak (10%, highest in the dataset)



Human-generated

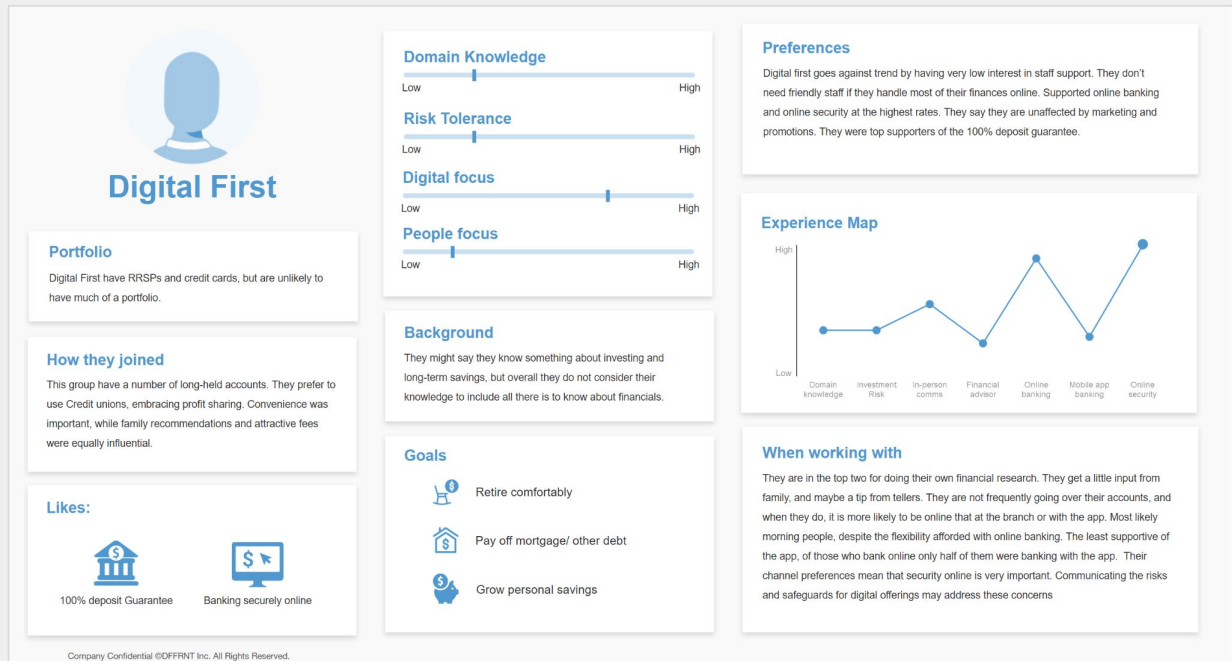
Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's



Claude-generated

Contraction: Digital First has "a number of long-held accounts" and "embrac[es] profit sharing." Priya is 1-4 years tenured and scores *lowest in the dataset* on profit share as a reason for banking (37%). Same channel behaviour, same security and deposit-guarantee priorities, opposite relationship to the institution and therefore opposite retention strategies



Human-generated

Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's

Cheryl is missing entirely from the human-generated persona set, and she's the largest Claude-generated segment at 18%.

Starters (human-generated) are low-knowledge but new and app-enthusiastic. In contrast, Cheryl is 11–20 years tenured, holds 2.4 products, barely uses the app (49%), uses no advisor, and barely uses the phone.

She's not at the start of a journey: she's been a member for fifteen years and the relationship has produced almost no product understanding.

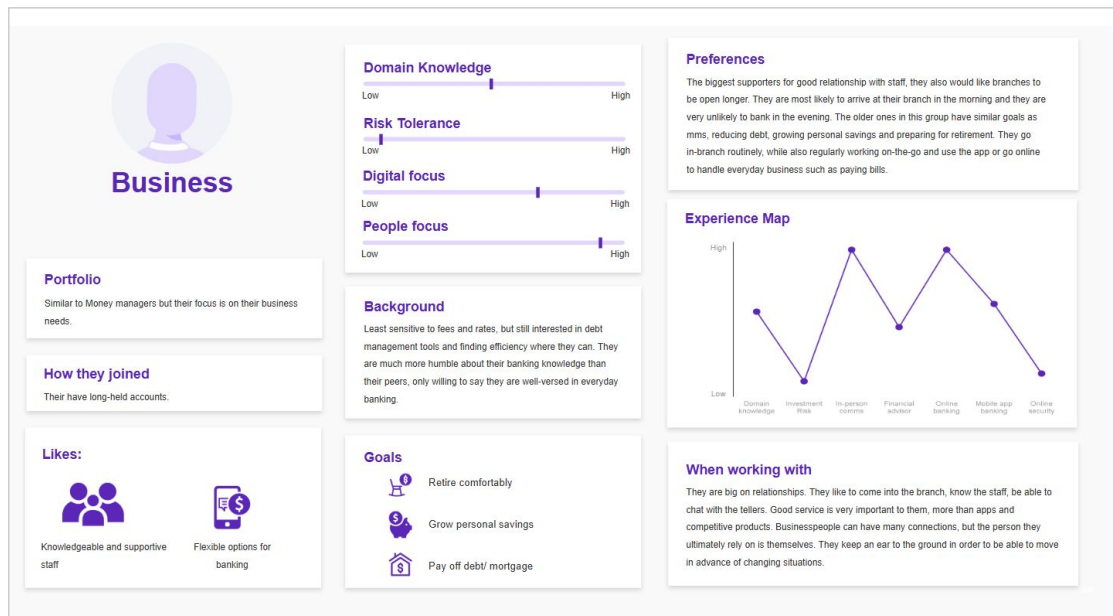
A new member with low confidence is a normal onboarding problem. A fifteen-year member with low confidence is an institutional failure, and it's a much bigger commercial opportunity.



Claude-generated, not identified in the human set

Appendix 1 - Synthetic Personas

Comparing Claude to human-designed Persona's



Human-generated, not replicated by the Claude analysis

Claude had no equivalent to the human-generated Business persona. Business banking shows up in the Claude personas (Dwayne holds a business account; Tyler & Jasmine are preparing to open one) but distributed across two personas rather than concentrated into its own.

Overall: Claude did an adequate job with the data it was given. The human-generated personas benefitted from in-depth scoping with the client, as well as interviews to further refine the persona categories.

- 2 Claude personas mapped very closely onto the human work
- 2 Claude personas were partial fits
- 1 Claude persona contradicted the human equivalent on key traits
- 1 Claude persona had no equivalent in the human set
- 1 human persona had no Claude equivalent

Bottom-line: Claude provided useful insights, quickly, but couldn't replicate the depth of real client engagement and follow-up interviews.

Appendix 2: Synthetic Usability Test

How did the simulation compare to real people?

15 synthetic participants (Sept 2026) vs DFFRNT's moderated test with 6 people and an expert site audit (Sept 2024)

81% vs 71%

task success,
synthetic vs human

-0.2

rank correlation of
task success

0.79

rank correlation of
task time

17 of 26

human findings
surfaced – only 9
of them fully

Two studies, two methods

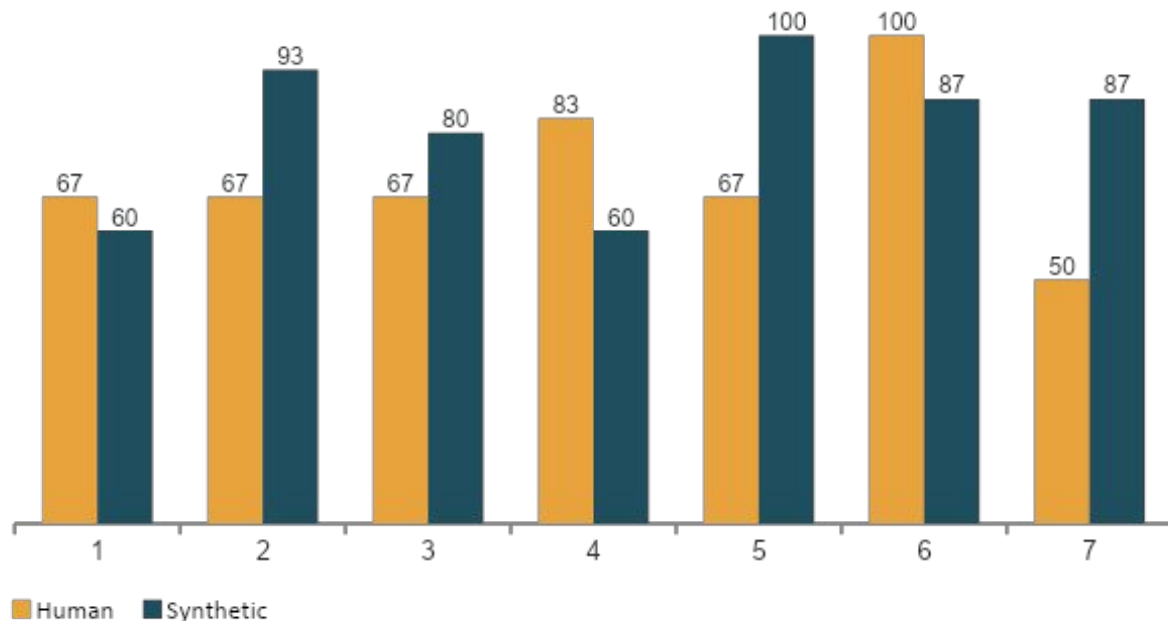
Not a like-for-like comparison – read the differences with these gaps in mind

	Human study (2024)	Synthetic study (2026)
Participants	6 people recruited via the City contact centre; mixed ages, genders, professions	15 model-simulated participants, one persona (middle-aged, calm, satisficing)
Session	Moderated, think-aloud, probing, post-task questions	Unmoderated simulation; think-aloud simulated in some runs
Site version	saskatoon.ca, 2024	Live saskatoon.ca, Sept 2026 (Bylaw 9990 in force; zoning pages reorganised)
Site search	Avoided by protocol (0 searches)	Allowed – used on 52% of attempts
Time on task	Measured; ~5-minute limit	Modelled estimate
Task wording	Original	Day care reworded; 'plan' → 'application' (portal); 'PDF' added (Rosewood)
Task order	Map, condo, Rosewood, bylaw, portal, stats, day care	Map, day care, stats, condo, portal, bylaw, Rosewood
Analysis	2 usability experts; site audit; accessibility scan	Coded from simulation output; scoring harmonised; key facts verified live

Some differences on the following slides come from these gaps (site changes, search, wording, order), not from the method itself.

Task success doesn't line up

% success by task – human (n=6) vs synthetic (n=15)



-0.2

rank correlation across the 7 tasks –
effectively none

~20 pts

average gap per task

+33 / +27

too optimistic: permit portal, condo

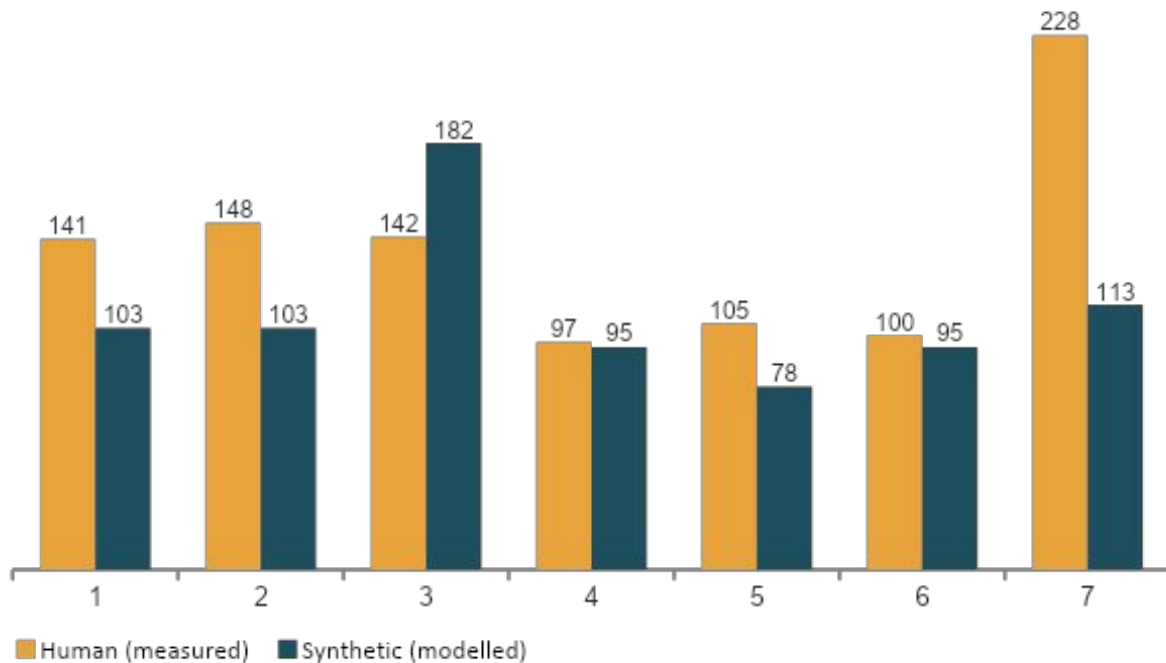
-23

too pessimistic: digital bylaw

*Day care wording differs between studies. Overall: synthetic 81% (85/105) vs human 71% (30/42). Small samples – directional only.

Time ranking tracked better than success

Human geometric mean (measured) vs synthetic mean (modelled), seconds



0.79

rank correlation of task time – the order of difficulty was broadly right

Rosewood

overstated: 182s modelled vs 142s real (search dead ends humans never hit)

Day care

understated: 113s vs 228s – the hardest real task (wording differs)

17 of 26 human findings surfaced - only 9 full

Each issue in the human report, checked against the synthetic analysis

9 Matched	8 Partly	8 Missed	1
• Zoning content scattered • Too many competing maps • Duplicate pages • 'Discretionary Use' jargon • Day care page built on definitions • Concept Plan naming inconsistent • ePermitting slow and not understood • Outdated 2022 data • Business & Development overloaded	• Overlapping topics across pages • DZB links not clearly clickable • 'How to' links go to guides • Sub-sites: Saskatoon Land, DZB • Map links buried below text • Permit Forms: no context or filters • Concept Plans deep / below fold • <i>(named, but not the mechanism)</i>	• Mega menu too big – #1 human finding • Menu closes on hover • PDFs with no navigation • Left-side nav missing on guides • Alerts ignored (banner blindness) • Missing alt-text • Saskatoon Land page design • Wanting local zoning alerts	•

Contradicted (1): humans said Condominium Approvals needs a checklist; the simulation held it up as the model page.

Where the two studies agree

Same issue, independent evidence

Finding	Human evidence (2024)	Synthetic evidence (2026)
Zoning content scattered	<i>"I need to do a lot of clicking around to find this..."</i>	3–4 zoning pages; which one you land on decides the bylaw task
Competing maps	<i>"I think this is the map I need but I know there's another one..."</i>	12/15 opened the Zoning Address Map; 6 stopped there
Bylaw link buried	<i>Link "embedded in a paragraph" rated MAJOR</i>	Only link text is "here"; one run read it twice and missed it
Business & Development is a catch-all	<i>"Business and Development feels like a catch-all..."</i>	10/15 looked in City Hall or Community first for statistics
Permit Forms pulls people in	<i>"I see Permit Forms and see the button to apply, I think I could apply here"</i>	8/15 clicked or were drawn to Permit Forms & Applications
Jargon	<i>"Discretionary use" is passive jargon; "what is digital?"</i>	"Discretionary Use" read off the page, never used unprompted; "digital" read as "online" in all 6 bylaw failures
Home-based licence trap	<i>One participant ended on the home-based business licence</i>	Home Based page misled 2/15

What the simulation missed, added, and got wrong

Added

- Mega menu: too long, too many columns, closes on hover – the #1 human finding
- Permit Forms filter box and 'Apply' label – caused a time-out
- Sub-sites (Saskatoon Land, DZB) trapping users
- Map usability: legends, zooming, load time
- Right-hand links unnoticed on wide screens
- Attitudes: “my default is to phone”; wanting zoning alerts

Missed

- Search behaviour – excluded from the human protocol (still unverified)
- Order effects quantified – the same risk existed in the human study: ePermitting seen during the condo task; the DZB found during Task 1
- Explicit scoring rules and a data-quality audit

Got wrong

- Called Condominium Approvals a model page – humans want a checklist
- Rated stale statistics High – humans called it minor, and one found the newer report
- Called the portal the easiest task (100%) – humans: 67%, with a time-out and a wrong answer
- Invented site facts – 6 runs described the ZAM as interactive

Where synthetic users fit

Use it for

- Pressure-testing tasks, wording and order before fieldwork
- Hypotheses about labels, IA and content
- Catching order effects and scoring gaps
- Cheap re-checks after content fixes

Don't use it for

- Success rates, times or benchmarks
- Menus, hover, scrolling, filters, maps
- Attitudes, trust and real-world context
- Facts about the site without verifying them

