

Heura Eval Report 004: Arche 0.19 vs 0.20 vs 0.21

Oct 14, 2026 · @Ryan

Arche 0.21 is much better on the business benchmarks and about the same as 0.20 in the Studio. It more than doubles both business benchmarks (v1 89.2%, v2 68.3%, before the code checks) and keeps nearly all of 0.20's learning from taught answers (GEN-1 45/68). It practised these 16 task types, so the checks still waiting on one download (the code checks, BIZ-FRESH and CODE-1) will show how far the jump carries.

In the Studio it gains the most from training (Trailhead's exam 62% to 88%, level with 0.20) and scores 2 of 6 on the two hard customer questions before and after training (0.20: 3, then 1). One Fix still carries over (S1b 3/3). Ryan's Verity gets 4 of 9 on 0.21, retrained or not; live today she gets 2.

Scorecard

0.21 is far ahead on both business benchmarks and within one case of 0.20 everywhere else.

Measure	Arche 0.19	Arche 0.20	Arche 0.21	0.21 vs 0.20	Notes
Business AI Benchmark v1 (20 tasks)	29.2%	33.3%	89.2%	Better	Mean of 3 runs, before the code checks. With them: 0.19 27.5%, 0.20 33.3%; 0.21's aren't scored yet.
Business Extension v2 (20 tasks)	not run	32.8%	68.3%	Better	Content checks, 3 runs. Its sandboxed code checks aren't scored yet.
GEN-1: learn from one taught answer (68)	32/68	46/68	45/68	Within 1 case	65 of 68 cases scored as on 0.20: 1 gained, 2 lost.
Picks the right tool (30 tasks)	30/30	30/30	29/30	Within 1 case	Opened an image link instead of viewing it. The base model alone: 9/30.
Studio check, unseen hotel (10)	10/10	10/10	10/10	Same	Cites the doc, admits gaps, follows a rule.

Measure	Arche 0.19	Arche 0.20	Arche 0.21	0.21 vs 0.20	Notes
Trailhead exam, untrained (8)	88%	75%	62%	Within 1 case	Two answers are true but incomplete (no standard shipping for a P.O. box, no 30 days to add Care), so they count as misses.
Trailhead exam, after training (8)	75%	88%	88%	Same	0.21 gained the most: two more questions right. 0.19's drop was kept off: v0 stayed live.
S1 + S2 before training (of 6)	2/6	3/6	2/6	Within 1 case	S1 calls a day-40 tent outside the return window; S2 ignores the manager.
S1 + S2 after training (of 6)	2/6	1/6	2/6	Within 1 case	S1 pastes the kit's gift-card example but invents nothing; 0.20 invented a policy.
S1b after one Fix (of 3)	3/3	3/3	3/3	Same	Also gives the Canada delivery time (the customer lives in Toronto).
Verity, untrained on this base (9)	not run	4/9	4/9	Same	The same 4 right; the sequence now copies the taught example.
Verity, trained on this base (9)	2/9	3/9	4/9	Within 1 case	0.19 = her live v2, re-run today. 0.21 got peppermint right, hummingbird wrong.

The Studio rows for 0.19 and 0.20 are Eval 003's runs in the new Studio (Oct 13); 0.21 ran the same steps today. Benchmarks for 0.20 and 0.21 come from the Kaggle run of Oct 14 (0.20 scored the same on the Mac on Oct 13), 0.19's from Eval 003 and its own benchmark report.

What changed in 0.21

0.21 is 0.20 trained one more epoch on new business and coding rows, plus a replay of 0.20's data.

- **Trained on Kaggle** (two T4 GPUs) from 0.20: 5,713 rows, 1 epoch, 714 steps, learning rate $1e-4$. Loss fell from about 0.80 to under 0.1. The run, with its benchmarks, took 2 hours.

- **1,600 rows of business work:** the 16 task types of the two business benchmarks (complaints, campaigns, sales, expenses, launches, policies, invoices, reorders, ticket routing, approvals and more), 100 new variants each, half of them advanced. Every answer comes from a solver that also reproduces both benchmarks' answer keys; no row copies a benchmark case.
- **1,520 coding rows** whose code passed its own tests in a sandbox, 895 MBPP problems, and 1,698 rows replayed from 0.20's training so it keeps what it had.
- **Installed in Heura Studios** on Oct 14 at 10:26. 0.20 is kept to switch back.

What we saw

In the Studio, 0.21 often stops at the first half of an answer, and twice it copied a taught example instead of answering.

- **Half answers in the Studio.** Two of its untrained exam answers are true but incomplete, so they count as misses. "Sorry, no. Express shipping isn't available to P.O. boxes." never says standard shipping works, and its Care answer drops the 30 days to add it.
- **S1 says no outright.** "The tent has been over the 30-day window ... so it's outside our return window." The Refund Policy gives store credit from day 31 to 60, and a snapped pole belongs to the warranty, which 0.21 never mentions.
- **S2 is kinder but less useful.** "I'm sorry for the frustration. I don't have that information." It is the only S2 that apologizes, but it never answers the manager request; 0.19 and untrained 0.20 said plainly they couldn't connect one.
- **It copies taught examples.** After training, its S1 answer is the kit's gift-card example nearly word for word. On Verity's sequence puzzle it repeats the taught example's numbers and that example's answer, 206 (the right answer is 136).
- **Training and the Fix still work.** The kit's 10 examples lifted its exam from 62% to 88%, and v1 went live. One Fix carried over to S1b with every detail traceable.

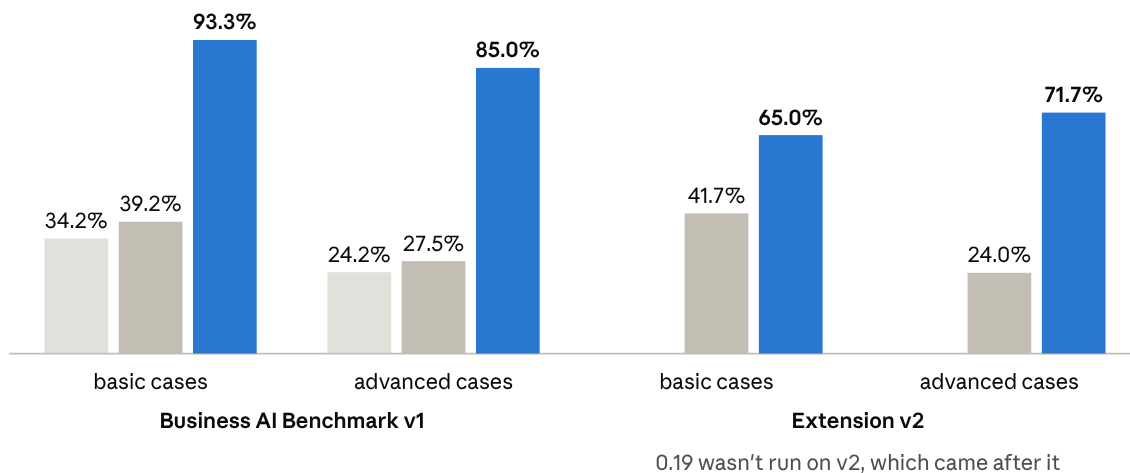
Business benchmarks

0.21 more than doubles both business benchmarks overall, and on the advanced cases of both it roughly triples 0.20's score.

0.21 more than doubles both business benchmarks overall

Content checks passed by case level, mean of 3 runs, before the code checks

Arche 0.19 Arche 0.20 Arche 0.21



Kaggle run of Oct 14 (0.20, 0.21) and the 0.19 benchmark report of Oct 12 · 20 cases per benchmark, mean of 3 runs

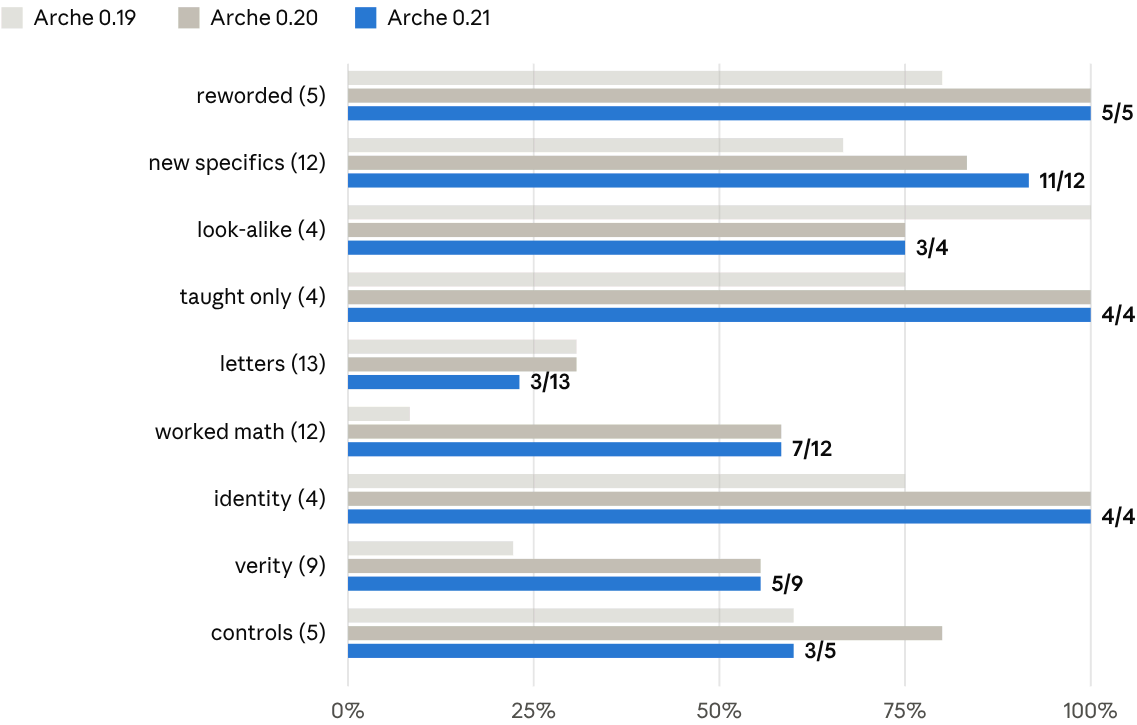
0.21 practised all 16 task types with new facts, so these scores measure skill on practised types. Its code checks aren't scored yet (0.19 and 0.20 passed none of v1's four). BIZ-FRESH's 36 new cases of the same types will show whether the jump holds on new facts, and CODE-1's unpractised part and HumanEval whether it carries to new kinds of tasks.

GEN-1: learning from one taught answer

0.21 kept nearly all of what 0.20 learned: 65 of the 68 cases scored the same, and every group is within one case. GEN-1 asks about companies, words and puzzle numbers never seen in training, each with one taught example in the Studio brief.

GEN-1: 0.21 stays within one case of 0.20 in every group

Share of cases right in each group; the number of cases is in brackets



gen_check.json for 0.19 and 0.20 (Oct 13) and the Kaggle run's log for 0.21 (Oct 14) · 68 cases in 9 groups

Question	Arche 0.19	Arche 0.20	Arche 0.21
can I pause my membership while I'm travelling?	✗ "I don't have that information. Please contact the Summit Fitness team..."	✓ "Yes. You can freeze it for up to 3 months a year at \$5 per month."	✓ the same
can I leave my cat with you while I'm on vacation?	✗ "I don't have that information. Please call the clinic..."	✓ "Yes, we board cats and small dogs for \$40 a night."	✓ the same
Halloween with 23 children (new numbers)	✗ 19 pieces	✓ 4 pieces, Ryan's method	✓ 4 pieces, Ryan's method
how many 's's are there in sassafras	✗ "s-a-s-s-a-f-r-a-s. That's five"	✓ four	✓ four

Question	Arche 0.19	Arche 0.20	Arche 0.21
what time does the pool close on saturday?	✓ "one hour before the club, which is 7 pm on Saturday"	✗ "On Saturday we close at 8 pm."	✓ "one hour before the club, so at 7 pm on weekends"
how many 'f's are there in coffeeshouse	✓ two	✓ two, spelled "c-o-f-f-e-e-h-u-s-e"	✗ "c-o-[f]-[f]-e-e-[f]-e-[h]-u-s. That's three"
price floor: \$18 to make, at least a 40% gross margin	✗ 7200	✓ 30	✗ 27
how many 's's are there in possessive	✗ two	✗ one	✗ one, spelled "p-o-p-e-l-[s]-i-v-e"

The first four rows are 0.20's gains that 0.21 kept. The next three are the only cases that changed from 0.20 to 0.21; the price floor was a 0.20 gain that 0.21 lost. Possessive fails on every version.

The Studio protocol: Trailhead

0.21 gained the most from training and scored 2 of 6 on the two hard questions before and after it, within one point of 0.20. The protocol is Eval 001's, run in a throwaway Studio (the live one at localhost:8000 was never touched): a new Trailhead from the Customer Support kit for Northwind Outfitters, S1 and S2, the kit's exam, Train on the kit's 10 examples, the same again, then one Fix on S1 and S1b. Claude scored the answers with Eval 001's rubric: C1 the right policy or source, C2 no invented facts or promised actions, C3 every part covered. As in Eval 003, a real rule applied to the wrong case costs C1, not C2.

Step	0.19 (Eval 003)	0.20 (Eval 003)	0.21 (today)
Exam questions missed, untrained	45-day store credit	45-day store credit; Care's 30 days	45-day store credit; P.O. box (standard shipping); Care's 30 days
Train on the kit's 10 examples	82 rows, loss 0.45 to 0.34, 6 min	82 rows, loss 0.37 to 0.18, 6 min	82 rows, loss 0.38 to 0.17, 7 min

Step	0.19 (Eval 003)	0.20 (Eval 003)	0.21 (today)
What went live	v0 kept (v1 scored 75%)	v1 at 88%	v1 at 88%
S1 before → after training (of 3)	0 → 0	1 → 0	1 → 1
S2 before → after training (of 3)	2 → 2	2 → 1	1 → 1

0.21 on	Its answer	Score
S1, before training	"Unfortunately not. Personal care items can't be returned, and the gift card is one of the items we don't accept. The tent has been over the 30-day window and isn't marked Final sale, so it's outside our return window."	1/3. Nothing invented, but the wrong rules: the return path instead of the warranty, and day 40 is inside the store-credit window.
S1, after training	"Unfortunately not. Personal care items, for hygiene reasons, can't be returned. The same goes for clearance items marked Final sale, personalized or engraved items, and opened personal care items."	1/3. The kit's gift-card example: real rules, the wrong case.
S2, before training	"I can't answer that. I don't have information about that. I'm sorry for the frustration."	1/3. Nothing invented; the manager request goes unanswered.
S2, after training	"I'm sorry for the frustration. I don't have that information."	1/3. The same.

0.21 on	Its answer	Score
S1b, after one Fix	"It's been 5 weeks, which is past our 30-day return window, so a return isn't available. But a snapped zipper is a defect, so it's covered by the warranty instead. Open the warranty claim form ... We reply within 2 business days. Since you're in Canada, international delivery takes 7 to 14 business days."	3/3. Every detail traces to the Fix and the kit's examples.

S1 fails on every version for the reason Evals 001 to 003 found: its brief gets the Shipping FAQ and the Refund Policy, never the Warranty doc. The 45-day store-credit question fails on every version too. The Fix text's "a return isn't available" is the protocol's own slip (store credit runs to day 60), and every S1b repeats it.

Ryan's Verity

On 0.21, Verity gets 4 of the 9 right whether she is retrained or not; her live v2 gets 2. These are the same 9 held-out versions of what Ryan asked her, run on a copy of her folder (her live versions were not changed). The 0.20 column is Eval 003's run of Oct 13; she has one more example now.

Question (right answer)	Live v2 (trained on 0.19)	0.20, untrained	0.21, untrained	0.21, retrained
s's in mississippi (4)	loops: m-i-s-s-i-s-s-i-s-s...	five	five	five
p's in peppermint (3)	loops: p-e-p-e-p-e...	four (misspells it)	four	three ✓
letters in hummingbird (11)	13	11 ✓	11 ✓	15
Halloween, new numbers (8)	13	8 ✓, Ryan's method	8 ✓, Ryan's method	8 ✓, Ryan's method

Question (right answer)	Live v2 (trained on 0.19)	0.20, untrained	0.21, untrained	0.21, retrained
Sequence 39, 54, 82 (136)	133	131	206, the taught example's answer	134
who am i talking to	Verity ✓	Verity ✓	Verity ✓	Verity ✓
what do you think about chatgpt	short, fair ✓	fair ✓	fair ✓	fair ✓
render distance, can't find it	"I'm not able to help with that."	Ryan's thumbs-up answer (vague)	the same	the same
how do i make a furnace	nonsense	wrong (charcoal + wood)	wrong (charcoal + glass)	wrong

Retraining on 0.21 used 84 rows, 4 of her lessons new since v2. It didn't hurt her overall, as it did on 0.20 (3 of 9), but the letters still swing: peppermint is now right and hummingbird went to 15. Her letter lessons still count without marking the letters.

Next

1. **Score the rest of the Kaggle run.** bench-answers.zip (100 KB) holds both versions' answers for v1's code checks, v2's sandboxed code, CODE-1 and BIZ-FRESH. BIZ-FRESH shows whether the jump holds on new facts. Downloading the file needs Ryan's OK. HumanEval failed to run for both; its log is in the run's output.
2. **Keep the Studio voice in 0.22.** Add Studio rows from other companies whose answer needs its second half: the alternative, the middle tier of a rule (like days 31 to 60), the escalation. Never the kit's own tests.
3. **Fix retrieval for long questions.** Open since Eval 001: search each sentence of the question, or let the model pick the documents, so the Warranty doc reaches S1's brief.
4. **Verity is Ryan's call.** On 0.21 she gets 4 of 9 untrained or retrained, against 2 today. Reteach her letter lessons in the marked style before the next retrain.

Limits of this run

- Greedy decoding, one run per case for GEN-1, the protocol and Verity: differences of 1 or 2 cases are noise, and each exam question is worth 12.5 points.
- 0.21's benchmarks ran on Kaggle, 0.19's on the Mac. 0.20 ran in both places and scored the same (GEN-1 46/68, v1 33.3% before the code checks).
- The business numbers are before the code checks, on task types 0.21 practised with new facts.
- Claude scored the protocol and Verity answers, as in Evals 001 to 003. The close call: 0.21's S1 states wrong rules ("outside our return window") without inventing any, so it keeps C2, as 0.20's did in Eval 003. Read strictly, 0.21 would score 1 of 6 before and after training.
- 0.21's protocol ran on today's Studio. For Heura Arche, its brief, example matching and Train button are unchanged since Eval 003.
- Verity has one more example than on Oct 13, so live v2's answers moved on 4 questions; her score didn't.