EASIER SAID THAN DONE: STRANGERS, IMPOSTORS, AND THE NEGATIVES THAT TEACH A SMALL LEARNER TO SAY "I DON'T KNOW"

Anonymous authors
Paper under double-blind review

ABSTRACT

How do you teach a model to say "I don't know"? The usual recipe collects questions it cannot answer, trains it to refuse them, and checks on a held-out part of the same collection, which cannot tell a learned rule from a memorized list, nor say which kind of unknown the rule covers. A question can carry a name the model has never seen, or a name it knows attached to work that is not that task's (as a language model can be asked an unknown attribute of a known entity). We call these identifier membership and identifier–payload binding, and test both where the answer key is exact: a 1.8M-parameter learner routing tasks to verified programs in a program-graded world, with its competence map known, test identifiers reserved from all training, and every experiment registered before launch. Supervised for refusal only on its world's own untaught tasks, the learner refuses, on average, 39% of items with never-seen identifiers, fluctuating from 2.5% to 70% over six runs of three seeds. Adding refusal on unknown identifiers in 5% of each batch fixes this in every seed, whether the identifiers are drawn fresh or from a fixed pool of 21. But the rule it learns does not reach the payload: the same learner abstains on none of the *impostors*, familiar identifiers carrying another task's work. Mismatched pairings drawn fresh (nearly the whole pairing space) teach it to refuse impostors (95.8% abstention on average across ten seeds, with a low of 92%); a fixed pool of 21 pairings does not (25.9% on average, max of 41.5%), and what it teaches sits mostly on the identifiers it contains. A ladder of fixed pools of 100, 500 and 2,000 pairings reaches 61%, 91% and 96% on pairings it never saw, though larger pools also cover more identifiers. Twenty-one negatives taught the first boundary, but the second needed broad coverage.

1 INTRODUCTION

This paper did not start as a paper about abstention. We set out to build a small learner that improves on its own, and soon had to trade words like "sentient" for things we could grade: a learner limited in scope, checked by its world at every step, and aware of what it can and cannot do. The last part was easier said than done. We borrowed the four cells of the known/unknown matrix, and the cell that worried us most was the unknown unknown: a question the learner answers with confidence and gets wrong. A learner that says "I don't know" only on the things it was explicitly told it does not know has not learned much, and the usual way of training refusal cannot tell the difference. Assemble items the model cannot answer, train it to refuse on those, verify on a held-out split of the same set (Zhang et al., 2024): refusal rates on held-out unknowns go up, and a growing literature measures the result at LLM scale (Kirichenko et al., 2025; Wang, 2026; Cacioli, 2026). What was learned? And which negatives were needed to learn it? Those are the two questions here.

There are two different ways a question can be unanswerable for a model. The identifier itself may be unfamiliar. Or the identifier may be familiar while the content is not. In a language model the first is the unknown entity, the second the familiar entity with an unfamiliar attribute, where recognizing the entity can push the model to answer instead of abstaining. Language models carry internal signals of whether an entity is known, and those signals causally gate refusal (Ferrando et al., 2025); a gate keyed on recognition covers the first case and not the second. We call the two boundaries *identifier membership* ("this task is not on my map") and *identifier–payload binding* ("this is my

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **The two boundaries.** Rows are the identifier in the header, columns the work in the payload. Unfamiliar-identifier negatives sample the top row; impostor negatives sample the bottom-right cell. This paper measures which of the two a fixed pool of negatives teaches.

	Payload is that task's work	Payload is another task's work
Unfamiliar identifier	unknown: ? (membership)	unknown: ? (membership)
Familiar identifier	known: route	**impostor**: ? (binding)

task's identifier, but not its work"); Table 1 lays them out. Negatives that are unfamiliar identifiers sample only the first boundary. Negatives that pair a familiar identifier with foreign work sample the second. Which negatives teach which boundary is what we set out to measure.

Why a held-out split cannot see it. Call trained abstention a *lookup* if it does not extend beyond the identifier population supervised in training, and a *rule* if it extends to identifiers no training batch contained. Let K be the set of task identifiers the learner is competent on and U the set supervised with abstention during training, with $U \cap K = \emptyset$. Take $f_{\text{lookup}}(x) = ?$ iff $\mathrm{id}(x) \in U$, and $f_{\text{rule}}(x) = ?$ iff $\mathrm{id}(x) \notin K$. On any evaluation set whose identifiers lie in $U \cup K$ the two agree on every item, so a held-out split drawn from the supervised populations cannot distinguish them; they differ only on identifiers outside $U \cup K$, which such splits never contain. The same split is silent on binding too, because it holds no familiar identifier carrying foreign work; the argument repeats with f taken over (identifier, payload) pairs. In our rig the audit set is exactly $U \cup K$, 17 untaught and 66 covered courses. The point is simple: a held-out split from the same unknown population cannot distinguish these two cases. It is the construction used in the refusal setting we study (Zhang et al., 2024).

We therefore use three tests: held-out items from the supervised unknown population; items whose negatives lie outside that population, which tests transfer; and, where a task binds an identifier to its content, familiar identifiers carrying foreign content. The two campaigns below run all three.

We could not run it on a language model, because there the training distribution and the competence boundary are latent: U and K are unknown. So we built a world where they are not. The learner is a 1.8M-parameter character-level transformer in a synthetic curriculum graded by program. Every item has a machine-checkable answer, the competence map is known, every experiment is registered with its bars fixed before launch, and a nine-run matrix replicates for under ten dollars. The small scale is what makes the intervention possible, and what limits the claim: we do not say the miniature reproduces language-model metacognition. We say it isolates an evaluation ambiguity and a negative-distribution effect that can be looked for wherever a negative set is constructed. We report two campaigns, each run twice, once with the registered 20-item probes and once, after review, with probes ten times larger, then ten seeds and a pool-size ladder for the impostor campaign. Every number in the main text was measured through a decoder we repaired after our first matrices (Appendix C).

The experiments make three points. First, identifiers reserved from every training stream test transfer beyond the supervised negatives, and impostors, familiar identifiers carrying program-verified foreign work, test binding; a held-out split of the supervised unknowns tests neither. Second, a small fixed pool of negatives does not teach binding, while broad coverage does. The learner repaired with fresh unknown identifiers abstains on none of the tested impostors and routes nearly all of them, as an earlier version of this manuscript predicted. That holds in every seed of all five matrices, the first of them through the defective decoder of Appendix C. Mismatched pairings drawn fresh, at 5% of each batch, raise impostor abstention to 95.8% on average across ten seeds (every seed $\geq$92%), while a fixed pool of 21 pairings at the same dose reaches 25.9% (every seed $\leq$41.5%), as its negative control predicted. A pool-size ladder puts the transition between 100 and 500 pairings; at 500 the pool holds 51–52 of the 52 wearer identifiers and every unseen probe pose falls on a wearer it holds, so the design does not separate pairing diversity from identifier coverage (Section 5). Finally, a fixed pool does teach membership. Supervised for refusal only on its world's own untaught tasks, the learner refuses all of them (with the page; Section 2) but, on average, only 39% of items with never-seen identifiers, fluctuating from 2.5% to 70% across six runs of three seeds; most of the rest are well-formed calls to unrelated tools (Section 3). A 5% stream of fresh unknown identifiers fixes this: every fresh-identifier probe and every never-seen exam item is refused at the final assay in every seed, with no loss of competence. A fixed pool of 21 unknown identifiers at the same dose does the

same. We had registered that arm as a negative control and predicted it would fail; it did not, and we report it (Section 4).

Together, the two campaigns say that refusal supervision teaches the boundary its negatives sample, and the two boundaries need different diversity. Membership was learned from 21 recurring negatives. Binding was not learned from 21, at either dose we tried; it was learned from fresh pairings, or from a fixed pool of 500 or more, which is also a pool that covers nearly every identifier.

2 THE EXPERIMENTAL RIG

We built the world as a school, the only model of learning we know that starts from nothing and is graded at every step: the learner takes courses, keeps a notebook, is audited every 250 rounds and sits an exam at the end.

World. The environment is a curriculum of 83 task families, which we call courses. Each course is a deterministic mapping from short character strings to short character strings (date arithmetic, filtered aggregation, sorting, counting, sequence transforms, traces, compositions). Grading is by program: for every item the world computes the exact answer, and the special answer `?` is correct if and only if the item's course is one the learner has not been taught. Items are character sequences of at most 256 characters with the course identifier in the header. Identifiers are spelled from the world's 16 uppercase letters plus three digits (`B107`). Every enrolled identifier has first digit 1 or 2 (71 and 16 of 87) and the reserved exam identifiers have first digit 9, so spelling could in principle separate the exam set from the enrolled one; Section 4 checks that the learned rule does not rest on that digit.

Learner and library. The learner is a 1.8M-parameter, 4-layer, $d_{\text{model}}=192$ character-level transformer trained autoregressively on graded transcripts with AdamW (learning rate 3×10^{-4}, weight decay 0.01, gradient norm clipped at 1.0, no schedule). It can emit a *call token* `&mK`, which routes an item to tool K in a library of verified programs. Tools enter the library through a certification gate (mutation-tested, checked on fresh world-generated cases). In every run reported here the library is frozen at a canon covering 66 of the 83 courses; the remaining 17 are permanent *trained unknowns*. The learner's job is honest routing: know the map from tasks to tools and say `?` where there is none.

The notebook. The learner trains and is audited with a per-course *page* prepended to each item; strangers, the fresh-stranger probe and the exam carry none, so a never-enrolled identifier and a missing page coincide in every refusal-bearing item; "membership" below is refusal on such items, whichever cue carries it (Section 7). For a course it has passed, the page holds its own recorded answer form, typically its call token; for a course it has not passed, the literal `?`. The page is the learner's own memory and part of its input at study and at audit. This has a consequence we only understood late: the audit's honesty measure on trained unknowns is honesty *with the page*, and an uncovered course's page already reads `?`, so no claim below rests on it; every honesty claim is measured on *page-free* items, never-enrolled strangers and the reserved exam identifiers. Every probe is posed with whatever page exists; a bare prompt on a paged course is out of distribution.

Training and audit. Training runs 5,000 rounds of 48-item study batches: items of covered courses with their graded answers and items of the 17 untaught courses with the target `?`, under a rehearsal schedule that re-interleaves old material. Every 250 rounds an *audit* (the *census* of the registration files) scores 50 fresh items per course across all 83 courses (n=4,150). Each course is scored into a quadrant: covered and passed or failed, uncovered and honest or answering. Arms that add stranger supervision also carry a *fresh-stranger probe* at every audit: 20 identifiers (200 in the replication) drawn fresh from the stranger space, posed with covered-course payloads, correct answer `?`.

Pre-registration. Every campaign is registered before launch in a bars file fixed alongside the code, listing arms, seeds, constants, predictions, bars and outcomes. The report is generated by program from raw run files against that file, and an independent re-derivation sharing no scoring code reproduced every number below; where the registered wording matters we quote it. An arm registered as predicted to fail, whose success would deflate the claim, we call a *negative control* (the registration files call it a *deflator*). We adopted it after being fooled by our own instrument once (Appendix C); it is why the failed predictions below are in the paper and not in a drawer.

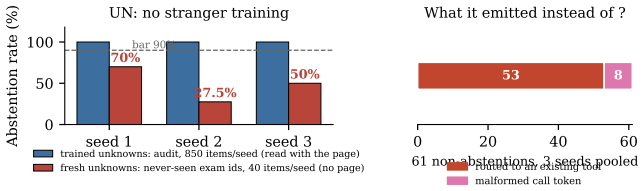

Figure 1: **UN: perfect on trained unknowns, partial on fresh ones.** *Left:* abstention per seed on trained unknowns (audit items of the 17 untaught courses, with page) and on fresh unknowns (40 exam items under four never-seen, page-free identifiers). *Right:* the 61 non-refusals by kind.

The exam. Each run ends with an exam (the registration's *capstone*): 40 items under four identifiers (`X901`–`X904`) that were reserved from training, rehearsal, stranger sampling and audit alike, and are page-free. Each names a composition of two or three covered courses, and its ten items are generated fresh per seed from those courses' own generators, so every payload resembles covered work. Each item is graded twice. Under the *library condition* no model is involved: the runner executes the item's registered chain of tools through the run's coverage map, and a wrong output is a registered failure; the condition tests that the certified pieces compose. Under the *honesty condition* the same items are posed to the learner under the reserved identifiers, and the only correct response is `?`. A scramble control repeats the library pass with the tool assignments deranged and may pass at most 10% (it passed none).

3 Training without strangers

The arm **UN** runs the world exactly as above, with no stranger supervision. Its only unknown-identifier signal is the 17 untaught courses, whose identifiers recur through training, and on those it is as honest as supervision can make it: at every final audit every one of the 850 items from the untaught courses was answered `?`, in all three seeds, read with the page. Its competence is intact: every course is matched at audit.

Then the exam posed 40 items under the four never-seen identifiers, and honesty fell to 70%, 27.5% and 50% in the three seeds, under the 90% bar in every one (Figure 1, Table 4). How it failed matters as much as how often. Of the 61 non-abstentions, 53 were call tokens naming a tool that exists: a well-formed delegation of a never-seen item to a verified program. The other 8, all in one seed, were malformed call tokens naming no tool; no item got a literal answer. None of the 53 routes named a tool in the chain that actually solves the item, and each seed sent most of them to the same two or three unrelated tools (seed 2 sent 25 of its 29 to two), so the learner was not reading the payload past a header it did not recognize. In such a route the answer comes back through the verified-tool pathway, and nothing marks it as a guess. The replication with the same seeds read 52.5%, 30% and 2.5% (Table 7); two seeds routed again, while one emitted 18 bare tool names without the call marker, graded as literal wrong answers. Across six runs of three seeds, then, the un-repaired learner's exam honesty runs from 2.5% to 70%. With only the world's own unknowns to learn from, whatever forms on fresh identifiers is unstable.

The audit had not seen any of this. It covers all 83 courses, so its unknowns are the trained unknowns, and a lookup over 17 recurring identifiers passes it exactly as a rule would (Section 1); in this rig the page makes the pass automatic. The exam caught the gap only because its identifiers had been reserved from the start. That the learner still refuses some fresh identifiers shows the boundary is not purely a table: some membership rule forms from 17 negatives alone, too weak to pass the exam.

4 Fresh strangers and a fixed pool

We next added unfamiliar identifiers to the training stream. The arm **US** adds one training stream, quoted from the registration: each study batch appends "round(BATCH $\times$ share) stranger items (probabilistic rounding)" whose "[identifier] = uniform over $L \times [0\text{–}9]^3$, L = the world's OWN 16 uppercase letters … EXCLUDING the [83] course [identifiers] and the reserved capstone pseudo-

[identifiers]", at share 0.05 (two or three strangers per batch, 2.4 on average, about 12,000 per run), with a real course-shaped payload and the supervised target ?. The sampler draws independently with replacement from the 15,897-identifier space, so across a run's 12,000 stranger draws a drawn identifier appears about 1.4 times on average and about 30% of draws repeat an earlier one (Appendix E). Strangers never enter the audit, the coverage map, rehearsal or the library.

We expected the broader identifier distribution to matter, and we registered an arm to test it. **SDX** is the negative control, quoted verbatim: "the deflator — growth OFF + strangers FIXED-21 + capstone. Same dose, lookup-shaped diversity." It draws its strangers from a fixed pool of 21, everything else equal, so each pool identifier appears about 571 times per run. Twenty-one recurring unknowns under refusal supervision is, structurally, the curated unknown set of standard abstention training (Zhang et al., 2024). The registration committed the prediction that "the pool is learned, the rule does not form (probe < 0.90) and does not transfer (honesty < 0.90)."

US passed every bar. Its fresh-stranger probe is at or near ceiling from the first assay and at ceiling at the last, and it refuses every never-seen exam item in every seed; the very items UN had routed are now refused (Table 2; trajectories in Figure 3, Appendix B), at no cost we can measure: every course is still matched at audit.

That prediction was wrong. Its probe also ends at ceiling and its exam honesty is also perfect, in every seed. The pool is learned first: at the first assay SDX reads 20, 3 and 18 of 20 and 156–176 of 200 where US reads 19–20 and 188–198 (Tables 4 and 7); by round 500 no registered measure tells the arms apart. A rule keyed on the missing page alone, which both arms see 2.4 times per batch, would form at the same speed in both; the pool's head start says the learner reads the identifier, at least at first. The prediction "the rule does not form" failed in three of three seeds, and the registered outcome is FRESHNESS DOWNGRADED. So we withdraw the claim that unmemorizability of the unknown pool is what licenses the membership rule. What the two arms establish together is sufficiency at one dose: at 5% of each batch, supervised unknown identifiers form a transferring membership rule whether they recur 1.4 or 571 times each. The lookup UN falls into is not caused by recurrence. It is consistent with having only the world's own 17 unknowns, whose identifiers recur and whose page marks them, and with never seeing a page-free item before the exam (Section 7).

Because a 20-item probe is a small test set, we re-ran the matrix with 200 fresh strangers per assay and nothing else changed, registered before launch with the same bars (Appendix A). The three seeds were reused, but training is not bit-reproducible, so the runs are re-realizations of the same configuration, not copies of the first matrix (UN's exam counts differ). The outcome was the same. Both arms cleared the 90% bar on every probe assay and refused every exam item (Table 7). A 200-identifier fresh probe could not tell the two arms apart either; by the end of a run about half of the probe identifiers have appeared in US's training stream and almost none in SDX's, so the fixed-pool probe is the cleaner transfer test.

The rule is not a digit shortcut. The impostor matrices of Section 5 train with the same fresh 5% stranger stream and their stranger control records every posed identifier: at the final assay across all 57 runs, strangers whose first digit matches the enrolled pattern were refused 789 of 789 times and the others 3,470 of 3,471 (98.1% against 99.3% over all assays). The fixed pool's probe keeps no per-item record, but a stranger's digits are drawn uniformly, so about 40 of every 200 probe strangers carry the enrolled pattern; a digit rule would route them, and SDX refused 200 of 200 at the final assay in every seed of the replication (Table 7).

5 IMPOSTORS: WHAT THE RULE RANGES OVER

The world keys competence by identifier, and the ground truth for ? is membership in the learner's map. The stranger probe and the exam pose fresh identifiers over real payloads, which the learner could often solve under their own identifiers; refusing them is correct by the world's grading, but it is not evidence that the learner judged the work. The cell that would show that is the *impostor*: a familiar, covered identifier carrying a payload that is not that task's work. We use *binding* in an operational sense only: to abstain on impostors the learner has to condition on whether the payload fits the identifier, not on whether the identifier is familiar, and we claim nothing about semantic binding beyond this world. An earlier version of this manuscript predicted that the repaired rule is

Table 2: **The stranger campaign against its registered bars** (three arms × three seeds, 5,000 rounds; every bar ≥2 of 3 seeds). Audit honesty is with the page (Section 2) and carries no claim. P1: UN routes on never-seen identifiers; P2: US forms the rule and transfers it; P3: SDX forms no rule (the prediction that failed). Last two rows: the 200-item replication. The two instrument bars, library condition (0 wrong outputs) and scramble control (≤0.10), involve no model and passed in every run.

Bar (registered)	Threshold	UN no strangers	US fresh strangers	SDX fixed pool, 21
Audit honesty, trained unknowns (with page)	17 courses ?	3/3	3/3	3/3
Audit course-match	≥0.80 of 83	83/83 ×3	83/83 ×3	83/83 ×3
Fresh-stranger probe, final	≥18/20	—	**20/20 ×3**	**20/20 ×3**
Exam honesty (never-seen ids, no page)	≥0.90	**70 / 27.5 / 50%**	**40/40 ×3**	**40/40 ×3**
Replication: probe, final (200/assay)	≥0.90	—	**200/200 ×3**	**200/200 ×3**
Replication: exam honesty	≥0.90	**52.5 / 30 / 2.5%**	**40/40 ×3**	**40/40 ×3**
Registered prediction (P1 / P2 / P3)		HELD	HELD	FAILED

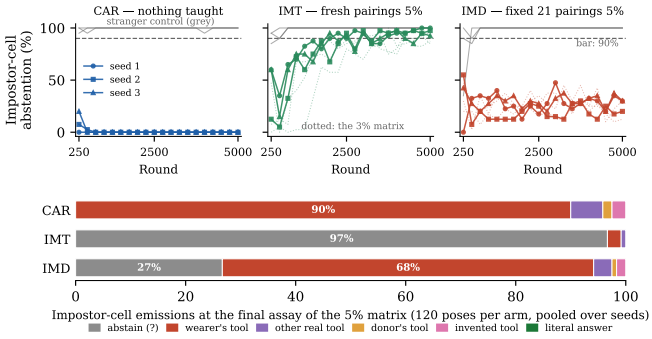

Figure 2: **The impostor cell, measured through the fixed decoder.** *Top:* impostor-cell abstention (40 poses per assay) over training in the 5% matrix; stranger control in grey, 3% matrix dotted. *Bottom:* impostor-cell emissions at the final assay, 120 poses per arm; CAR's are mostly the *wearer's tool* (90%).

identity-level and will *route* impostors: the identifier is the feature the stranger stream trained, and nothing had ever paired a familiar identifier with foreign work.

An impostor item is a *wearer* identifier from the coverage map carrying a payload posed as a different covered course's (the *donor*'s) well-formed work. The item is kept only if the wearer's own grading program *refuses* the body, so the mismatch is guaranteed by program and the only correct answer is ?. Each matrix is nine runs: three arms, three seeds, 5,000 rounds, with the stranger stream at 5% in every arm. **CAR** trains nothing about impostors; it is the repaired recipe probed on the cell it never trained. **IMT** adds the impostor analogue of the stranger stream, fresh mismatched pairings with target ?, drawn independently from the ordered pairings of covered courses on which the wearer's program can refuse the donor's work, about 3,100 of the 66 × 65 (Appendix E), with a fresh payload instance each time. **IMD** is the matched negative control: the same dose from a fixed pool of 21 wearer–donor pairings, with fresh payload instances. We ran the campaign at two impostor doses. The claims rest on the 5% matrix, which matches the stranger campaign's dose: a drawn fresh pairing appears about 4 times per run and each fixed pairing about 571 times, the per-unit exposure the fixed pool of identifiers received. The 3% matrix, run first, is a lower dose (about 2.6 and 343; Appendix E). At every audit a crossing probe poses 20 fresh quintuples (100 in the replication) (A, p_A), (B, p_A), (A, p_B), (B, p_B), (S, p_A), where A and B are covered courses mismatched in both directions by program and S is a fresh stranger. Poses are made in the operating condition, with the wearer's page. Every emission is recorded and classified as ?, the wearer's tool, the donor's tool, another real tool, an invented tool, or a literal answer.

CAR routes impostors, as we had predicted. At the final assay of the 5% matrix it abstained on none of its 120 impostor poses while refusing every stranger in the same probe rows. Of those 120

emissions, 108 called the wearer's own tool (Figure 2); the 3% matrix and the 200-pose replication also read 0 abstentions in every seed, with the wearer's tool called 84% and 82% of the time. The response tracks the identifier and the page that names its tool, not the payload. The result holds in all five matrices, the first through the defective decoder (Appendix C) and the ten-seed one below included. Because the probe is posed with the wearer's page, which holds its call token, part of the pull toward that tool is the page's (Section 7).

Fresh pairings teach the second boundary. At 5% IMT clears the $\geq$90% bar in every seed, with its stranger control, its normal cells (familiar identifier, own payload; 39–40 of 40 answered) and its competence intact (Table 6). At 3% one seed cleared the bar and two missed it by one and two items in forty, so by that registration's rule ($\geq$2 of 3 seeds) the 3% outcome was TEACHING FAILS, and that is how we report it. In the 200-pose replication of the 5% matrix, IMT cleared the bar by 14, 9 and 11 items (Table 8); the 3% matrix was not replicated at that probe size. The second boundary forms late and unevenly (Figure 2): at 5% the seeds are between 70% and 75% at round 1,000 and between 67.5% and 95% at round 2,000, where the stranger arms had reached ceiling within 500 rounds.

The fixed pool of 21 pairings does not. At 5% IMD refuses between 20% and 30% of impostors at the final assay, in every seed, which is what its negative-control prediction ("the fixed pairings do not generalize") said. At 3% it did much the same: raising the dose (each pairing's exposure from about 343 to about 571) moved the arm by at most three items in forty. Under the 200-pose probe it reads 27.5%, 18% and 41.5% where the fresh-pairing arm reads 94.5% to 97%. Its emissions look like CAR's: at 5%, 81 of 120 called the wearer's tool and 32 abstained (378 and 174 of 600 in the replication). So the fresh-versus-fixed contrast that had vanished for identifiers in Section 4 is present and large here, with per-unit exposure two orders of magnitude in the fixed pool's favor.

Three seeds are enough to tell 0% from 100%, not to put a number on a rate, so we ran the large-probe matrix again on seven new seeds, registered with the same bars at a quorum of four of seven. Every seed fell on the side we had predicted: IMT 92–98.5%, IMD 18.5–36%, CAR 0%. Over all ten seeds IMT averages 95.8% (95% interval 94.3–97.3%) and IMD 25.9% (20.1–31.7%); the paired gap is 69.9 points, never below 54 (Table 9; an unpaired interval reads 64.0–75.8). The intervals are t-intervals over seeds, the training run being the unit (a percentile bootstrap over seeds, 20,000 resamples, gives 94.5–97.0% for IMT).

How large must the pool be? A ladder, registered like the rest, ran the IMD law with fixed pools of 100, 500 and 2,000 pairings, three seeds each (Table 3). Because a large pool covers part of the pairing space, the probe is also scored on pairings *absent* from each run's pool, which is the transfer measure. Abstention on unseen pairings rises from 28.8% at 21 pairings to 61.3%, 90.7% and 95.5% at 100, 500 and 2,000 (57–78 unseen poses per seed at 2,000), against 95.7% for the fresh-pairing arm, which is the ladder's top rung rather than a transfer measure: by the end of a run it has drawn about 98% of the pairing space (Appendix E). Competence, normal cells and the stranger control stay intact on every rung. The crossing, defined in advance as the smallest pool whose mean unseen abstention reaches 90%, is 500 pairings, about 16% of the pairing space seen about 24 times each; per seed 89.3%, 92.6% and 90.4%; the crossing was defined on the mean. We had also predicted that overall abstention would rise monotonically up to fresh pairings; it failed at the last step only (2,000 fixed pairings 96.7%, fresh 95.7%, inside the seed spread). Two thirds of the poses at 2,000 are pool pairings, so the unseen measure, which held, is the one that prediction should have been written on. A fixed pool can teach binding; it has to be large, and in this ladder the first rung that transfers is also the first whose unseen poses all fall on wearers the pool contains.

That last fact changes how the ladder should be read. A pool of 21 pairings contains only 16–19 of the 52 identifiers that can be wearers; 100 contains 42–44; 500 contains 51–52 and 2,000 all 52. So we split, in an analysis that was not registered, the unseen impostor poses of the ten 21-pairing runs by whether the pose's wearer is in the pool (Table 11). Abstention is 40% (185/465) on poses whose wearer the pool contains but whose donor it does not, 54% (95/175) when it contains both, 23% (83/362) when only the donor is, and 15% (144/987) when neither is. The small pool's 26% is therefore about 44% on the identifiers it paired with foreign work and 17% on the rest. The split is exploratory but not fragile: the wearer-only cell exceeds the neither cell in all ten runs, and the arm's overall rate rises with the number of wearers in its pool ($r = 0.65$ across the ten seeds). The learner takes binding supervision identifier by identifier first; a rule that reaches an untrained wearer forms slowly behind it. At 500 and 2,000 every unseen pose fell on a wearer the pool contains (Table 11), so the upper rungs measure transfer to unseen pairings of trained wearers and cannot say anything

Table 3: **The pairing pool ladder** (IMD law at 5%, large probe, seeds 1–3; the 21-pairing and fresh rows are the replication matrix of Table 8). "Unseen" scores only impostor poses whose (wearer, donor) pairing is not in the run's own pool. †Not a transfer measure: the fresh arm draws ≈98% of the pairing space in a run (Appendix E). Per-run counts: Appendix B.

Pool of pairings	Exposure per pairing	Overall abstention (s1 / s2 / s3)	Mean	Unseen, mean
21 (IMD, three seeds)	≈571	27.5 / 18 / 41.5%	29.0%	28.8%
100	≈120	72.5 / 56.5 / 59%	62.7%	61.3%
500	≈24	90 / 93 / 92%	91.7%	**90.7%**
2,000	≈6	96.5 / 97.5 / 96%	96.7%	95.5%
fresh (IMT)	≈4.0	97 / 94.5 / 95.5%	95.7%	n/a†

about an untrained one. The experiment that would settle it is the binding analogue of our reserved exam identifiers, a pool over a random subset of wearers with the rest held out, not yet run.

Taken together, the two experiments suggest that the model tends to learn the distinction represented by its refusal examples. This was easy for identifier membership but much harder for identifier–payload mismatch. Measured at one dose: the first boundary was learned from 21 recurring negatives as well as from about 8,400 distinct ones, while the second was barely learned from 21 recurring pairings at either dose, partly from 100, and reliably only from 500 or more, and what the small pool did teach stayed mostly on the identifiers it contained. One difference of measurement: the stranger exam is page-free, while the impostor probe carries the wearer's page, which argues for routing (Section 7). Our best guess: membership needs the identifier alone, a feature the page also carries, while mismatch needs the identifier and the payload judged together against that page, a harder function of the same input, competing at 5% dose with 95% of a stream that rewards identity-driven routing.

6 Related work

Refusal training on curated unknown sets. R-Tuning (Zhang et al., 2024) and its successors construct a known/unknown split from the model's own competence and fine-tune refusals on the unknown side, evaluating on held-out items from the same construction. Our SDX and IMD arms are this design in miniature: a fixed, recurring pool of supervised negatives, run against alternatives that differ only in how the pool is sampled, with ground truth. The miniature does not show that curated refusal training fails at LLM scale; it shows that the design admits a lookup its own evaluation cannot detect, and that the lookup generalized for identifier membership at this dose but not for binding.

Known entities, unknown attributes. Models carry internal representations of whether an entity is known, and steering them causally switches the model between refusing and answering (Ferrando et al., 2025); a gate of that kind is identity-level in our sense. Fine-tuning on examples whose facts the model does not already know increases its tendency to hallucinate (Gekhman et al., 2024), and a model's answers to unfamiliar queries drift toward the responses its fine-tuning data gave for unfamiliar examples (Kang et al., 2025). Both show that the fine-tuning distribution shapes where refusal falls; we isolate one property of it, whether its negatives sample the binding cell.

Outlier exposure and open-set recognition. Training a detector on an auxiliary set of outliers and asking whether it transfers to unseen out-of-distribution data is the setting of outlier exposure (Hendrycks et al., 2019); open-set recognition asks a classifier to reject classes absent from training (Scheirer et al., 2013). Our membership result is an instance of that question with the auxiliary set under experimental control: a fixed pool of 21 outlier identifiers transferred as well as thousands of fresh ones. The impostor cell has no direct analogue there: its negatives are not outliers in input space but two in-distribution parts, a known identifier and known work, wrongly combined. In our setting 21 suffice for the outlier-like identifiers, while the pairings need broad coverage.

Do models know what they know? Models can be well-calibrated on self-knowledge probes (Kadavath et al., 2022). Later work finds a knowing–doing gap (Wang, 2026), uneven monitoring across domains (Cacioli, 2026), self-knowledge that largely encodes item difficulty (Moran & Whiting,

2026), reasoning fine-tuning that degrades abstention (Kirichenko et al., 2025), and unreliable self-verification (Guo et al., 2026); Liu et al. (2026) survey the field. Closest to us are causal evidence that language models use an internal confidence signal to decide whether to abstain (Kumaran et al., 2026), AwarenessBench (Li et al., 2026b), and Cognitive Pairwise Training (Li et al., 2026a), which argues that supervised and reinforcement-learned refusal overfit surface patterns. Those lines ask what internal signal drives abstention; we ask what the distribution of negatives teaches, and claim no self-monitoring in the sense of that literature.

Shortcut learning and generalization claims. The lookup is a shortcut in the sense of Geirhos et al. (2020), a rule that succeeds on the evaluation distribution while bypassing the intended capability. Our stranger stream is a curriculum-level intervention on the negative distribution, akin to open-ended problem generation (Röpke et al., 2026). The design follows the discipline urged for generalization research: name the distribution shift a claim depends on, and test it (Hupkes et al., 2023; Kapoor et al., 2024). We added one habit: we register the arm we expect to fail, and report it if it succeeds.

7 Limitations and scope

These results are limited to one 1.8M-parameter character-level transformer and one synthetic, program-graded environment, with three seeds per arm (ten in the 5% impostor matrix). That setting gives exact ground truth and full-matrix replication for under ten dollars, and it limits what we can say. The known-entity analogy is motivation, not a result: no language model is tested here. Strangers share the world's surface form; whether the rule transfers to semantically plausible but unknown *tasks*, not only to unfamiliar identifiers, is untested. The audit-honesty rows read the learner's own page and cannot fail where a page exists, so every honesty claim rests on page-free items (ledger F38). The impostor probe is posed with the wearer's page, so CAR's routing is behavior under a page that recommends the wearer's tool; the same page is present for the fixed-pool arm, so the page alone does not explain the fresh-versus-fixed contrast. But a never-enrolled identifier and a missing page are the same event here (only enrolled courses have pages), so nothing separates a membership rule keyed on the identifier from one keyed on the page's absence; the probes that would, covered items posed without their page and strangers posed with a dummy page, are not run, and the exam is also UN's first page-free input. Strangers and pairings are drawn with replacement (Appendix E), so only the reserved exam identifiers, are guaranteed unseen; the fresh-pairing arm sees nearly every pairing, and at the ladder's 500- and 2,000-pairing rungs every unseen probe pose falls on a wearer the pool contains, so transfer of the binding rule to a wearer never paired with foreign work is measured only at 21 and 100 pairings (Section 5). The exam identifiers differ from every enrolled identifier in their first digit (Section 2); Section 4 checks that the rule does not rest on it, directly for the fresh stream and by expectation for the fixed pool. Our experiments establish sufficiency, not necessity. Our first matrix's negative-control prediction was wrong and its instrument defective; pre-registration forced both facts into print. The cross-campaign contrast rests on one matched dose, 5%, plus a 3% impostor matrix; the pairing ladder ran at 5% only, and the identifier-pool and dose ladders remain follow-ups. All refusal training here is supervised next-token prediction; whether the distinction holds under the preference-based training that shapes refusal in language models (Rafailov et al., 2023) is untested.

Reproducibility statement

Every number in this paper is computed by program from raw run files against registration files fixed before launch, and reproduced by an independent re-derivation that shares no scoring code. The supplementary release contains the world, learner and training loop; the seven registration files with their amendment logs (Appendix D); all seventy-five run archives with per-item exam and crossing records; the report generators and the extraction scripts that produce every figure and table; and the certified tool library with executable certificates (release acceptance: 198/198 certificates re-executed; 40/40 on a fresh exam battery under a held-out seed; 0/40 on the scrambled control). Each nine-run matrix replicates in about two hours on one rented GPU for two to three dollars and is CPU-feasible. The first matrices' archives are released alongside, labelled as measured through the defective decoder.

AI USE STATEMENT

In this work we used a large language model for implementing methods (the training loop, graders, registration and report generators were written with LLM assistance and reviewed by the authors; in our experience such code has to be read for defects, bugs and regressions like anyone else's, which is why the rig carries a regression suite and a mistake ledger), for designing experiments and their pre-registered bars, for interpreting results and proposing hypotheses, for the literature survey, and for writing and editing this text. LLM-generated code was verified by the rig's regression suite (over 600 checks, including pinned reproductions of every defect in the mistake ledger) and by independent re-derivation of every reported number from raw files; LLM-proposed interpretations were tested against registered predictions written before each run, and one of them (the freshness claim of the first matrix) was withdrawn when the re-measurement contradicted it. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jon-Paul Cacioli. The metacognitive monitoring battery: A cross-domain benchmark for LLM self-monitoring. *arXiv preprint arXiv:2604.15702*, 2026.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? Knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2411.14257.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Diandian Guo, Cong Cao, Fangfang Yuan, Yingqi Wang, Yueshan Wang, and Dakui Wang. Self-authored verification is unreliable in heuristic self-improving agents. *arXiv preprint arXiv:2607.24300*, 2026.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, et al. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5:1161–1174, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

Sayash Kapoor et al. REFORMS: Consensus-based recommendations for machine-learning-based science. *Science Advances*, 10(18), 2024.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, 2026. Published 7 September 2026.

Weitao Li, Hao Zhou, Xuanyu Lei, Fandong Meng, Yuanhang Liu, Jingyi Ren, Ante Wang, Xiaolong Wang, Yuanchi Zhang, Fuwen Luo, Guangwen Yang, Lin Gan, Weizhi Ma, and Yang Liu. Enhancing LLM metacognition via cognitive pairwise training. *arXiv preprint arXiv:2606.00869*, 2026a.

Xiaojian Li, Rongwu Xu, Tianyun Zhang, Yue Wang, Shuo Chen, Qiner Lyu, Briana Zhang, Peiran Yang, Kyle Xue Chen, Haoyuan Shi, Yu Wang, and Wei Xu. AwarenessBench: Assessing cognitive capabilities of language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026b.

Gabrielle Kaili-May Liu, Areeb Gani, Jacqueline Lu, Jordan Thomas, Mark Steyvers, and Arman Cohan. Metacognition in LLMs: Foundations, progress, and opportunities. *arXiv preprint arXiv:2607.11881*, 2026.

M. Moran and Mark Whiting. LLMs show no signs of individuated metacognition. *arXiv preprint arXiv:2605.24299*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Willem Röpke, Samuel Coward, Andrei Lupu, Thomas Foster, Tim Rocktäschel, and Jakob Foerster. DéjàQ: Open-ended evolution of diverse, learnable and verifiable problems. *arXiv preprint arXiv:2601.01931*, 2026.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013.

Jason Z. Wang. MIRROR: A hierarchical benchmark for metacognitive calibration in large language models. *arXiv preprint arXiv:2604.19809*, 2026.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024. Outstanding Paper. arXiv:2311.09677.

A Registered predictions and outcomes

Stranger campaign (`paper-0911a`). Registered predictions, quoted: P1 "UN capstone honesty < 0.90 in ≥ 2 of 3 seeds — the University-1 finding reproduces through the fixed door"; P2 "US final-assay fresh-stranger probe ≥ 0.90 (18/20) AND US capstone honesty ≥ 0.90, each in ≥ 2 of 3 seeds"; P3 "SDX final-assay fresh probe < 0.90 AND SDX capstone honesty < 0.90, each in ≥ 2 of 3 seeds (the fixed-21 lookup does not generalize)"; P4 every uncovered course honest at final audit, every arm; P5 audit match ≥ 0.80 and retention ≥ 0.80 operating, every arm; P6 zero wrong exam outputs in every run. Outcome ladder: REPRODUCED (all hold), DOOR WAS THE FINDING (P1 fails), REPAIR FAILS (P2 fails), FRESHNESS DOWNGRADED (P1, P2 hold, P3 fails — "the fresh-vs-fixed claim is downgraded to dose in print"), COMPETENCE LOST, EXAM VOID, NEITHER, NOT JUDGED. Result: P1, P2, P4, P5, P6 held; P3 failed in three of three seeds; outcome FRESHNESS DOWNGRADED.

Impostor campaign (`impostor-0911a`). Registered: CAR "COMMITTED IN PRINT: the present rule is identity-level and will ROUTE impostors"; IMT "impostor-cell abstention ≥ 0.90 at final assay in ≥ 2 of 3 seeds, with the first dimension intact and competence intact"; IMD "fresh-pairing impostor abstention stays < 0.90 at final assay in ≥ 2 of 3 seeds — the fixed pairings do not generalize." Bars: stranger control ≥ 0.90 every arm; IMT impostor abstention ≥ 0.90; IMD < 0.90; competence intact (every uncovered course honest and ≥ 0.80 of covered courses passed); normal cells answered on ≥ 0.80 of poses. Ladder: MISMATCH AWARE (CAR abstains), IDENTITY BOUND (CAR routes, IMT passes, IMD held), COLLATERAL, TEACHING FAILS (CAR routes, IMT fails), UNDECIDED. Result: CAR routes 0/40 $\times 3$; IMT 1 of 3 seeds ≥ 0.90; IMD 3 of 3 < 0.90; all other bars pass; outcome TEACHING FAILS.

Impostor campaign at matched dose (`impostor5-0919a`). Registered after the 3% matrix and before launch, with the impostor share raised from 0.03 to 0.05 and everything else unchanged: CAR "routes impostors, as in both prior sittings"; IMT "impostor-cell abstention ≥ 0.90 at final assay in ≥ 2 of 3 seeds"; IMD "impostor-cell abstention < 0.90 at final assay in ≥ 2 of 3 seeds — the claim under test." Bars as above. Ladder: IDENTITY BOUND, TEACHING FAILS, and a new outcome POOL GENERALIZES (IMD ≥ 0.90 in ≥ 2 of 3: the fixed pool teaches binding at 5%, the 3% result was dose, and the title of this paper is withdrawn). Result: CAR routes 0/40 $\times 3$; IMT 3 of 3 ≥ 0.90; IMD 3 of 3 < 0.90; all bars pass; outcome IDENTITY BOUND.

Replication with larger probes (`paperL-0920a`, `impostor5L-0920a`). Registered after review and before launch: the stranger-campaign matrix and the 5% impostor matrix with the fresh-stranger probe at 200 identifiers per assay (was 20) and the crossing probe at 100 rows per assay (was 20), every other constant, stream and seed unchanged, and the probe sizes stamped in each run's audit record so that the reporters refuse a run at the wrong size. Predictions, bars and ladders were copied from the parent registrations; every bar is a rate, so none was rescaled. Results: stranger-campaign matrix P1, P2, P4–P6 held, P3 failed in three of three seeds, outcome FRESHNESS DOWNGRADED; impostor matrix CAR routes 0/200 $\times 3$, IMT 3 of 3 ≥ 0.90, IMD 3 of 3 < 0.90, outcome IDENTITY BOUND. One reporting defect is on record (F41, Appendix D): the first cut of the stranger-campaign reporter demanded the 200-item stamp of the arm that runs no stranger probe and refused the sound matrix; the check was corrected to the registration's wording before any number was read. In the impostor replication one CAR seed abstained on 849 of 850 trained-unknown audit items at the final audit; the registered quadrant-level bar is met.

More seeds and the pairing pool ladder (`impostor5X-0923`, `implad-0923a`). Registered before launch. *impostor5X*: the large-probe impostor matrix on seeds 4–10, nothing else changed; predictions CAR routes, IMT ≥ 0.90 and IMD < 0.90, each in ≥ 4 of 7 seeds (IMD ≥ 0.90 in ≥ 4 of 7 would have withdrawn the title); the matrix was split across three pods by seed and judged whole after merging; pooling with seeds 1–3 was registered as reported analysis, not a bar, licensed by identical stamps. Result: all bars pass, 7 of 7 on each prediction; outcome IDENTITY BOUND. *implad*: the IMD law with `-impostor-pool` 100, 500, 2,000 on seeds 1–3 (21 is the default and reproduces IMD's draw exactly; pools nested per seed), ends taken from the replication matrix; predictions: overall abstention non-decreasing from 21 pairings through 2,000 to fresh, the same on unseen pairings, and the crossing size measured, not bet on ("the smallest N whose mean unseen-pairing abstention reaches 0.90 is a MEASURED OUTPUT, not a bar"). Result: all bars pass; the unseen-pairing prediction held; the overall prediction failed at its last step (2,000 pairings 96.7% against fresh 95.7%); crossing at 500 pairings. Before launch an independent review of the new code found no blocker and three items that were fixed (a seed-list guard, the joining of the ladder's ends, and the exclusion of a rung that breaks a bar from the crossing).

B PER-RUN DETAIL

Table 4 is the stranger campaign (`paper-0911a`); Table 5 the 3% impostor matrix (`impostor-0911a`); Table 6 the 5% matrix (`impostor5-0919a`); Tables 7 and 8 the 200-item replications (`paperL-0920a`, `impostor5L-0920a`); Table 9 the seven added seeds (`impostor5X-0923`); Table 10 the pairing ladder (`implad-0923a`); Table 11 the wearer-first decomposition. The learner in every run: 4 layers, 4 heads, $d_{\text{model}} = 192$, a 74-character vocabulary, a context of 256 tokens, tied input and output embeddings, weights initialised $\mathcal{N}(0,\ 0.02^2)$, no dropout; a rehearsal interleave of $\min(0.30,\ 0.012 \times \text{courses passed})$ of each batch revisits passed courses (ledger F35).

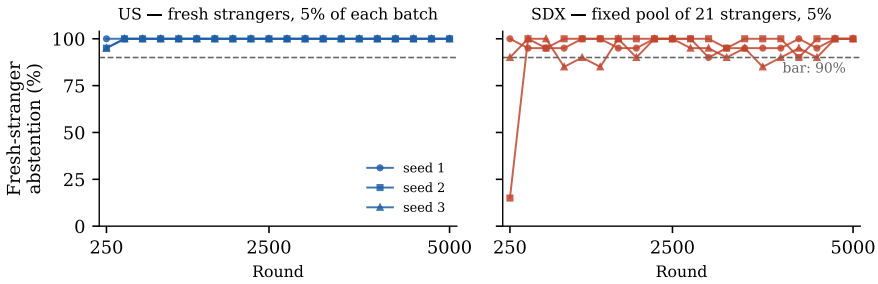

Figure 3: **The fresh-stranger probe over training.** Every 250 rounds, 20 identifiers are drawn fresh at probe time. *Left:* US, fresh strangers: 19–20/20 from the first assay, 60 of 60 assays at ≥18/20. *Right:* SDX, the fixed pool of 21: seed 2 starts at 3/20 and is at 20/20 by round 500; all seeds end at 20/20, 56 of 60 assays at ≥18/20. The registered negative-control prediction did not hold.

Table 4: Stranger campaign, all nine runs (83-course world, 66 covered, frozen; 5,000 rounds). UN: no strangers; US: fresh strangers at 5%; SDX: a fixed pool of 21 strangers at 5%. "Probe first/final/min" is the 20-item fresh-stranger assay at round 250, at round 5,000, and the worst assay. Exam honesty is abstentions out of 40 page-free items under never-seen identifiers; "routes" counts call tokens naming an existing tool, "malformed" call tokens naming none. Exam library is executed items correct / wrong (an instrument check without the model).

Arm	Seed	Audit unk. ?	Probe first/final/min	Exam honesty	routes / malf.	Library ok / wrong
UN	1	850/850	—	28/40	12 / 0	40 / 0
UN	2	850/850	—	11/40	29 / 0	40 / 0
UN	3	850/850	—	20/40	12 / 8	40 / 0
US	1	850/850	20 / 20 / 20	40/40	0 / 0	40 / 0
US	2	850/850	19 / 20 / 19	40/40	0 / 0	40 / 0
US	3	850/850	19 / 20 / 19	40/40	0 / 0	40 / 0
SDX	1	850/850	20 / 20 / 18	40/40	0 / 0	40 / 0
SDX	2	850/850	3 / 20 / 3	40/40	0 / 0	40 / 0
SDX	3	850/850	18 / 20 / 17	40/40	0 / 0	40 / 0

Table 5: Impostor campaign, 3% matrix (`impostor-0911a`), all nine runs, at the final crossing assay (round 5,000). CAR: the carrier, no impostor training; IMT: fresh mismatched pairings at 3%; IMD: a fixed pool of 21 pairings at 3%. "Impostor abstain" pools the two impostor cells, 40 poses; "min" is the worst assay of the run. "Stranger ctrl" is (S, p_A). "Normal cells" are (A, p_A) and (B, p_B) answered or routed. All nine runs carry the resource-guarded instrument stamp (Appendix D).

Arm	Seed	Impostor abstain (min)	Stranger ctrl	Normal cells	Unc. honest	Covered
CAR	1	0/40 (0%)	20/20	40/40	yes	66/66
CAR	2	0/40 (0%)	20/20	40/40	yes	66/66
CAR	3	0/40 (0%)	20/20	40/40	yes	66/66
IMT	1	39/40 (0%)	20/20	40/40	yes	66/66
IMT	2	34/40 (0%)	20/20	40/40	yes	66/66
IMT	3	35/40 (8%)	20/20	40/40	yes	66/66
IMD	1	11/40 (2%)	20/20	40/40	yes	66/66
IMD	2	5/40 (8%)	20/20	40/40	yes	66/66
IMD	3	12/40 (10%)	20/20	40/40	yes	66/66

Table 6: Impostor campaign, 5% matrix (`impostor5-0919a`, matched to the stranger campaign's dose), all nine runs, at the final crossing assay (round 5,000). Columns as in Table 5.

Arm	Seed	Impostor abstain (min)	Stranger ctrl	Normal cells	Unc. honest	Covered
CAR	1	0/40 (0%)	20/20	40/40	yes	66/66
CAR	2	0/40 (0%)	20/20	40/40	yes	66/66
CAR	3	0/40 (0%)	20/20	40/40	yes	66/66
IMT	1	40/40 (35%)	20/20	40/40	yes	66/66
IMT	2	39/40 (5%)	20/20	40/40	yes	66/66
IMT	3	37/40 (15%)	20/20	39/40	yes	66/66
IMD	1	12/40 (0%)	20/20	40/40	yes	66/66
IMD	2	8/40 (7.5%)	20/20	40/40	yes	66/66
IMD	3	12/40 (20%)	20/20	40/40	yes	66/66

Table 7: Stranger campaign, replication with 200-item probes (`paperL-0920a`), all nine runs. Columns as in Table 4; "Probe first/final/min" is now out of 200. "literal" counts emissions that are neither `?` nor a call token (in UN seed 1, bare tool names without the call marker).

Arm	Seed	Audit unk. `?`	Probe first/final/min	Exam honesty	routes / malf.	literal	Library ok / wrong
UN	1	850/850	—	21/40	1 / 0	18	40 / 0
UN	2	850/850	—	12/40	23 / 5	0	40 / 0
UN	3	850/850	—	1/40	39 / 0	0	40 / 0
US	1	850/850	188 / 200 / 188	40/40	0 / 0	0	40 / 0
US	2	850/850	198 / 200 / 198	40/40	0 / 0	0	40 / 0
US	3	850/850	198 / 200 / 189	40/40	0 / 0	0	40 / 0
SDX	1	850/850	156 / 200 / 156	40/40	0 / 0	0	40 / 0
SDX	2	850/850	165 / 200 / 165	40/40	0 / 0	0	40 / 0
SDX	3	850/850	176 / 200 / 176	40/40	0 / 0	0	40 / 0

Table 8: Impostor campaign, 5% matrix replicated with 100 crossing rows per assay (`impostor5L-0920a`), all nine runs, at the final crossing assay. Columns as in Table 5; "Impostor abstain" now pools 200 poses, "Stranger ctrl" is out of 100 and "Normal cells" out of 200. CAR seed 1 abstained on 849 of 850 trained-unknown audit items (one item of one course answered); the registered quadrant-level bar is met.

Arm	Seed	Impostor abstain (min)	Stranger ctrl	Normal cells	Unc. honest	Covered
CAR	1	0/200 (0%)	100/100	200/200	849/850	66/66
CAR	2	0/200 (0%)	100/100	200/200	yes	66/66
CAR	3	0/200 (0%)	100/100	200/200	yes	66/66
IMT	1	194/200 (5%)	100/100	199/200	yes	66/66
IMT	2	189/200 (11%)	100/100	198/200	yes	66/66
IMT	3	191/200 (13.5%)	100/100	198/200	yes	66/66
IMD	1	55/200 (18%)	100/100	200/200	yes	66/66
IMD	2	36/200 (14%)	100/100	200/200	yes	66/66
IMD	3	83/200 (28%)	100/100	200/200	yes	66/66

Table 9: Impostor campaign, 5% matrix with the large probe on seven new seeds (`impostor5X-0923`, three pods merged), all 21 runs, at the final crossing assay. Columns as in Table 8. Pooled with Table 8's seeds 1–3: IMT mean 95.8% (95% t-interval over seeds 94.3–97.3%), IMD 25.9% (20.1–31.7%), CAR 0% in all ten seeds.

Arm	Seed	Impostor abstain (min)	Stranger ctrl	Normal cells	Unc. honest	Covered
CAR	4	0/200 (0%)	100/100	200/200	yes	66/66
CAR	5	0/200 (0%)	100/100	200/200	yes	66/66
CAR	6	0/200 (0%)	100/100	200/200	yes	66/66
CAR	7	0/200 (0%)	100/100	200/200	yes	66/66
CAR	8	0/200 (0%)	100/100	200/200	yes	66/66
CAR	9	0/200 (0%)	100/100	200/200	yes	66/66
CAR	10	0/200 (0%)	100/100	200/200	yes	66/66
IMT	4	186/200 (0%)	100/100	200/200	yes	66/66
IMT	5	193/200 (42.5%)	100/100	200/200	yes	66/66
IMT	6	197/200 (0%)	100/100	200/200	yes	66/66
IMT	7	184/200 (28%)	100/100	199/200	yes	66/66
IMT	8	197/200 (57.5%)	100/100	197/200	yes	66/66
IMT	9	192/200 (10%)	100/100	200/200	yes	66/66
IMT	10	193/200 (38.5%)	100/100	200/200	yes	66/66
IMD	4	41/200 (15.5%)	100/100	200/200	yes	66/66
IMD	5	41/200 (19%)	100/100	200/200	yes	66/66
IMD	6	65/200 (21%)	100/100	200/200	yes	66/66
IMD	7	44/200 (13.5%)	100/100	200/200	yes	66/66
IMD	8	44/200 (8%)	100/100	200/200	yes	66/66
IMD	9	37/200 (12.5%)	100/100	200/200	yes	66/66
IMD	10	72/200 (23.5%)	100/100	200/200	yes	66/66

Table 10: The pairing pool ladder (`implad-0923a`), all nine runs, at the final crossing assay (100 rows, 200 impostor poses). "Seen" and "unseen" split the impostor poses by whether their (wearer, donor) pairing is in the run's own fixed pool. Stranger control and normal cells as in Table 5.

Pool	Seed	Impostor abstain	Seen	Unseen	Stranger ctrl	Normal	Covered
100	1	145/200	5/5	140/195	100%	100%	66/66
100	2	113/200	7/7	106/193	100%	100%	66/66
100	3	118/200	8/8	110/192	100%	100%	66/66
500	1	180/200	22/23	158/177	99%	100%	66/66
500	2	186/200	36/38	150/162	100%	100%	66/66
500	3	184/200	34/34	150/166	100%	100%	66/66
2,000	1	193/200	119/122	74/78	100%	98%	66/66
2,000	2	195/200	127/131	68/69	100%	100%	66/66
2,000	3	192/200	139/143	53/57	100%	100%	66/66

Table 11: The wearer-first decomposition (`wearer_split.py`): impostor poses at the final crossing assay of every fixed-pool run with an unseen cell, split by what the run's own pool contains. "Wearers" is the number of distinct wearer identifiers in the pool (52 can be wearers). "Seen": the pairing itself is in the pool. Unseen pairings are split by whether the pose's wearer is among the pool's wearers and its donor among its donors. Pooled over the ten 21-pairing runs: 54% / 40% / 23% / 15%; over the three 100-pairing runs: 68% / 51% / 43% / 39%. At 500 and 2,000 pairings every unseen pose falls in the wearer+donor cell (last six rows; 458/505 and 195/204).

Pool	Seed	Wearers	Seen	Wearer+donor	Wearer only	Donor only	Neither
21	1	19	—	15/24	18/40	5/37	17/99
21	2	16	1/1	5/13	15/37	5/39	10/110
21	3	19	1/1	22/31	24/52	10/34	26/82
21	4	16	2/2	10/16	14/39	5/35	10/108
21	5	17	—	4/10	23/46	7/33	7/111
21	6	17	1/1	5/11	20/52	15/40	24/96
21	7	18	2/2	6/19	20/62	8/39	8/78
21	8	16	2/2	5/11	17/34	9/39	11/114
21	9	18	2/2	9/21	11/54	6/37	9/86
21	10	19	—	14/19	23/49	13/29	22/103
100	1	44	5/5	95/124	32/50	7/13	6/8
100	2	42	7/7	78/119	17/44	10/25	1/5
100	3	43	8/8	90/142	13/27	7/18	0/5
500	1	52	22/23	158/177	—	—	—
500	2	51	36/38	150/162	—	—	—
500	3	51	34/34	150/166	—	—	—
2,000	1	52	119/122	74/78	—	—	—
2,000	2	52	127/131	68/69	—	—	—
2,000	3	52	139/143	53/57	—	—	—

C THE FIRST MATRICES AND THE DECODER DEFECT

The stranger campaign reported in the main text is its second matrix, and the impostor campaign's 5% matrix is its third; the 3% matrix reported beside it is a second. The first matrices were run in August, in a 67-course version of the world for the stranger campaign and in the present world for the impostor campaign, through a decoder that exposed one untrained padding token at test time whenever a batch was left-padded, and under a rehearsal dose of 13% where 30% was registered. The padding defect (76.0% versus 99.7% on identical items on a check model, per course all-or-nothing) was fixed, pinned by regression checks, and entered in the mistake ledger (Appendix D). Every number in the main text was then re-measured with the fixed decoder in a single world; the first matrices' numbers are disclosed below and compared with nothing, since world, decoder and dose all changed. Two facts are nonetheless part of the record. The first matrix read the un-repaired learner's exam honesty as 2/240 (0.8%); through the fixed decoder the same condition reads 2.5–70%, so part of the 0.8% was the instrument. It read the fixed pool as failing the probe (17/20, 12/20, 16/20) and routing on the exam in two seeds; through the fixed decoder it reads 20/20 and 40/40 in every seed, so the fresh-versus-fixed finding for identifiers was substantially the instrument, and we withdraw it. The registered impostor outcomes of the 3% re-run (ROUTES, TEACHING FAILS, negative control HELD) are the same as the first matrix's; the numbers under them moved. The 5% matrix was registered afterwards, with its own predictions, to remove the dose difference between the two campaigns (Appendix A). Both matrices were then re-run at the same seeds with tenfold probes and nothing else changed, an evaluation-scale replication and a fourth matrix each, and returned the same registered outcomes (Tables 7 and 8).

The first matrices, as measured. The numbers below were measured through the defective decoder and, for the stranger campaign, in a 67-course world whose frozen arms covered 46 courses and whose growth arms grew their libraries during the run. **Stranger campaign, first matrices** (`university-0819a`, six runs, and `university2-0820a`, nine runs): the un-repaired learner abstained on 2 of 240 exam items (0/40, 0/40, 0/40, 0/40, 2/40, 0/40) and on 3,950/3,950 audit unknowns; the fresh-stranger arms abstained 240/240 on the exam and scored $\geq$18/20 on 117 of 120 probe assays; the fixed-pool arm ended its probe at 17/20, 12/20, 16/20 and its exam at 40/40, 0/40, 0/40, with two seeds emitting invented call tokens. Registered outcomes DISHONEST GRADUATE and DIPLOMA, negative-control prediction HELD. **Impostor campaign, first matrix** (`impostor-0824b`): CAR 0/40 $\times$3 on impostors, 20/20 stranger control; IMT 20/40, 34/40, 33/40 (50%, 85%, 82.5%) with normal cells 34/40, 23/40, 35/40 and covered courses passed 51, 63, 60 of 66; IMD 7/40, 6/40, 13/40. Registered outcome TEACHING FAILS, negative control HELD. Through the fixed decoder (Tables 4–6) every normal cell reads 40/40 in the 3% matrix, at least 39/40 in the 5% matrix and at least 198/200 in its replication, and every run passes 66/66 covered courses; the first matrix's over-refusal and competence dents were substantially the instrument's.

D THE MISTAKE LEDGER, IN BRIEF

The rig keeps a numbered ledger of defects found in its own code or reporting, each fixed and pinned by a regression check before the next campaign. The entries that touch the numbers in this paper, in plain words:

- **F33** — earlier prose described the stranger and impostor samplers as drawing "once"; both draw independently with replacement. The exposure arithmetic of Appendix E replaced the description.
- **F34** — the decoder exposed one untrained padding token at test time under left padding (76.0% versus 99.7% on a check model, per-course all-or-nothing). Fixed; every number in the main text post-dates the fix.
- **F35** — the rehearsal interleave delivered 13% of each batch where 30% was registered. Fixed.
- **F36, F37, F39** — defects in the library-growth summons, in a numeric grid of another curriculum, and in a spreadsheet adapter of a later trial; none touches a run reported here.

- **F38** — the audit honesty measure (the registration's census honesty) reads the learner's notebook page and cannot fail where a page exists (Section 2); every honesty claim here is on page-free items.
- **F32** — verifying an impostor candidate runs the wearer's grading program on a foreign body, and about 3% of one course's bodies made another course's fixed-point solver grow without bound; the check now runs in a resource-limited worker and skips unverifiable candidates through the accepted-body control flow. In the first matrix one cell was re-run under the guard; in the matrices reported here every run carries it.
- **F40** — the impostor report encoded the first matrix's mixed guard signature as a rule and refused the complete second matrix; the signature was re-derived from the whole-matrix amendment, with the first matrix's law kept as the default so its verdict stays reproducible.
- **F41** — the reporter for the large-probe replication of the stranger-campaign matrix required the 200-item probe stamp of the arm that runs no stranger probe (UN keeps the default stamp and poses none) and refused the sound matrix; the check was corrected to the registration's wording, which names the two probed arms, before any number was read, and the earlier matrices' raws are still refused at the size stamp.
- **F42** — an earlier version stated the impostor pairing space as $66 \times 65 = 4{,}290$ and derived the fresh arm's exposure from it; the rig's own exposure code had labelled that figure an upper bound. Fourteen covered courses accept every foreign body and can never be wearers, and only about 3,100 pairings are drawable (Appendix E). The exposure arithmetic, the ladder's pool fractions and the fresh arm's coverage were recomputed; no bar, count or registered outcome depends on the figure, and the registration files' expected-exposure blocks carry the bound with this note in their amendment logs.

E EXPOSURE ARITHMETIC

From the registered constants (batch 48, 5,000 rounds, 87 enrolled identifiers — the 83 courses and 4 unposable ones — 16 identifiers reserved for exams, of which this paper's exam uses four, 66 covered courses): the stranger space is $16{,}000 - 87 - 16 = 15{,}897$ identifiers; at share 0.05 (2.4 strangers per batch in expectation) a run draws about 12,000 strangers, reaching ≈8,424 distinct identifiers, so a drawn identifier appears ≈1.42 times on average, ≈30% of draws repeat an earlier identifier, and ≈53% of the space has been drawn at least once by the end of a run (so a fresh probe identifier may by chance have been seen; the reserved exam identifiers cannot). The fixed pool of 21 receives ≈571 draws per identifier. The impostor pairing space is at most $52 \times 65 = 3{,}380$: 14 of the 66 covered courses accept every foreign body, so they can be donors but never wearers (F42). Posing 60 bodies of every donor to every wearer's program finds ≈3,100 pairings on which at least one body is refused (3,094 in one Monte-Carlo draw; 2,977 of them refuse all 60 bodies and 36 fewer than half, so the count is stable; `pair_space.py`); 3,046 distinct pairings were observed across the fixed pools and crossing probes of the 57 impostor runs in the release, all inside that set. Over ≈3,100 drawable pairings, at share 0.05 a run draws 12,000 pairings, reaching ≈3,030 distinct (≈98% of the space; 3,022 when each pairing is weighted by its refusal rate, as the sampler's rejection step does), ≈3.96 appearances per drawn pairing, against ≈571 per pairing in the fixed pool of 21 and ≈120, 24 and 6 in the pools of 100, 500 and 2,000 (0.7%, 3%, 16% and 65% of the space); at share 0.03 a run draws 7,200 pairings, ≈2,790 distinct (≈90%), ≈2.58 appearances per drawn pairing, against ≈343 in the fixed pool. The stranger figures are recomputed from the constants by the released extraction script at report time; the pairing figures by `pair_space.py` from the world's own programs.