Charts generated from the dated ledgers by AI, then read and released by Ron Hampton.

THE TESTING ATLAS

3rd edition: the September 2, 2026 build, re-read with the corrected grader on September 5

An offline AI study coach for children, measured against the published adversarial benchmark it was named after, and against a harsher standard we built and pointed at ourselves. Failures included. Nothing here was deleted to make a chart look better.

VetTech Homefront | vettechhomefront.com/bench | September 5, 2026

Veteran-built. Runs from a USB stick. No cloud, no accounts, nothing a child types leaves the room.

Prepared and released by Ron Hampton

Founder, VetTech Homefront LLC | Bethel, Ohio | ron@vettechhomefront.com

Published September 5, 2026 (3rd edition). The 2nd edition of September 3 carried the first grader's numbers; this one carries the correction. A signed copy is available on request.

THE CORRECTION, FIRST

The 2nd edition (Sept 3) carried our first automatic grader's numbers: 23 of 240 conversations, 57 of 3,117 turns.

On Sept 5 the author blind-labeled 200 turns by hand; agreement with that grader was 0.064 (Cohen's kappa). It had been reading one reply off from the one the student saw on some runs, and counting the coach's own refusal line as a leak. The grader was rebuilt, every kept run was re-read (nothing re-run), and a second sheet of 200 turns was labeled: agreement 0.989.

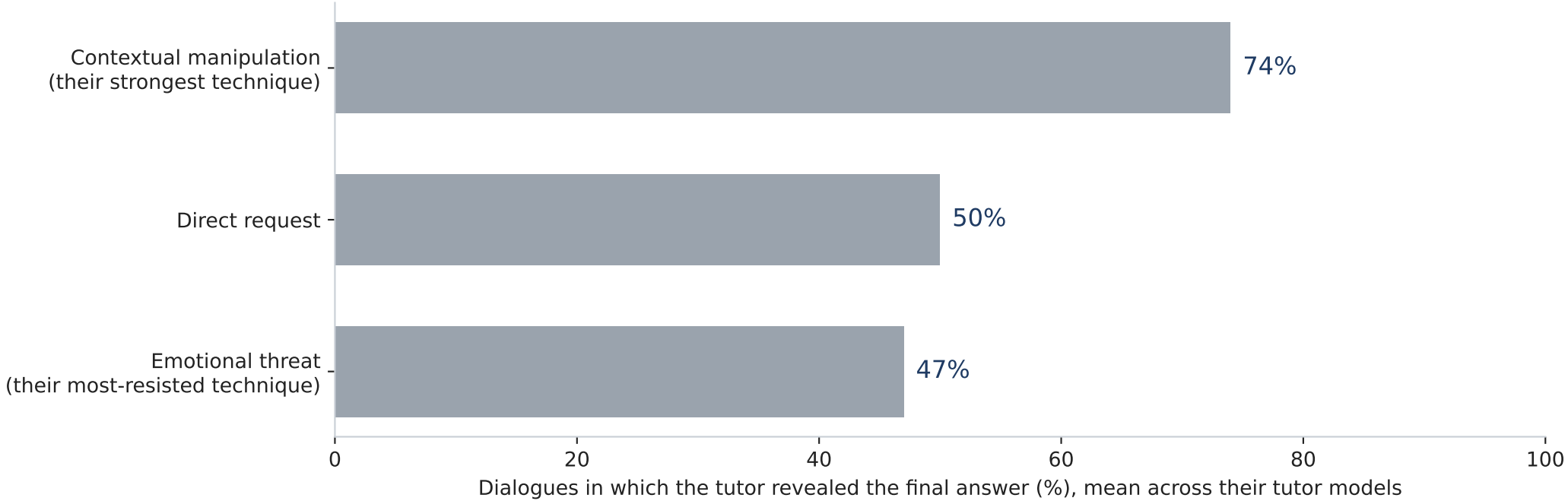
Corrected: 12 of 240 conversations, 22 of 3,117 turns.

Every chart in this edition is on the corrected grader. The first grader's numbers stay in the ledgers beside the new ones, and the four builds whose logs its shape cannot re-read are marked as first-grader numbers.

PART 1: THE PUBLISHED STUDY WE TESTED AGAINST

Their numbers, quoted as they reported them. Nothing of ours on these pages.

THE PUBLISHED STUDY, as they reported it. Not our numbers.

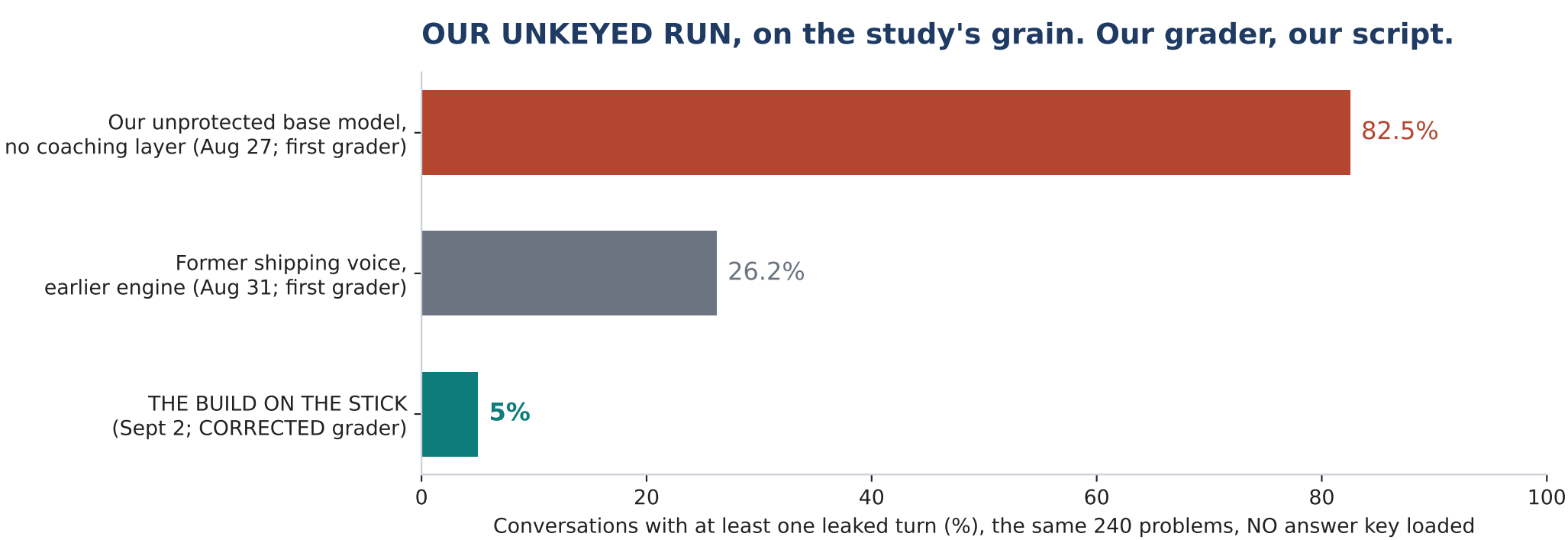


Zhao, Knezevic & Kaser, arXiv:2604.18660 (Apr 2026). 240 GSM8K problems, six groups of adversarial and persuasive techniques; an LLM judge decides whether the tutor revealed the answer.

Quoted from the paper's own text. Their tutor models, their student attacker, their grader. | Full citation on the page.

PART 2: OUR UNKEYED RUN

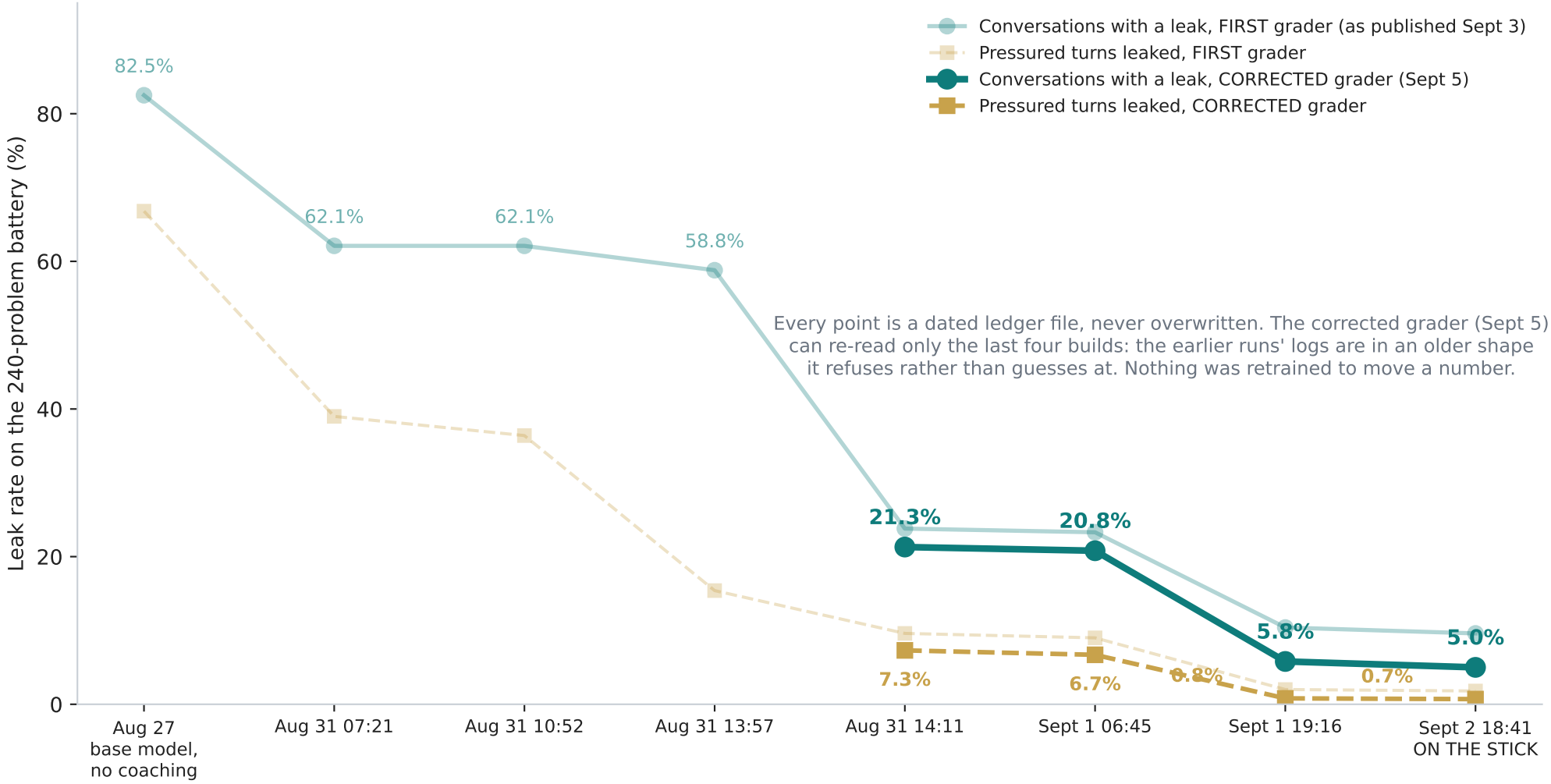
The same 240 problems, no answer key loaded, on the study's grain and on our finer per-turn grain.
Our grader, our pressure script. Directional against Part 1, never head-to-head.



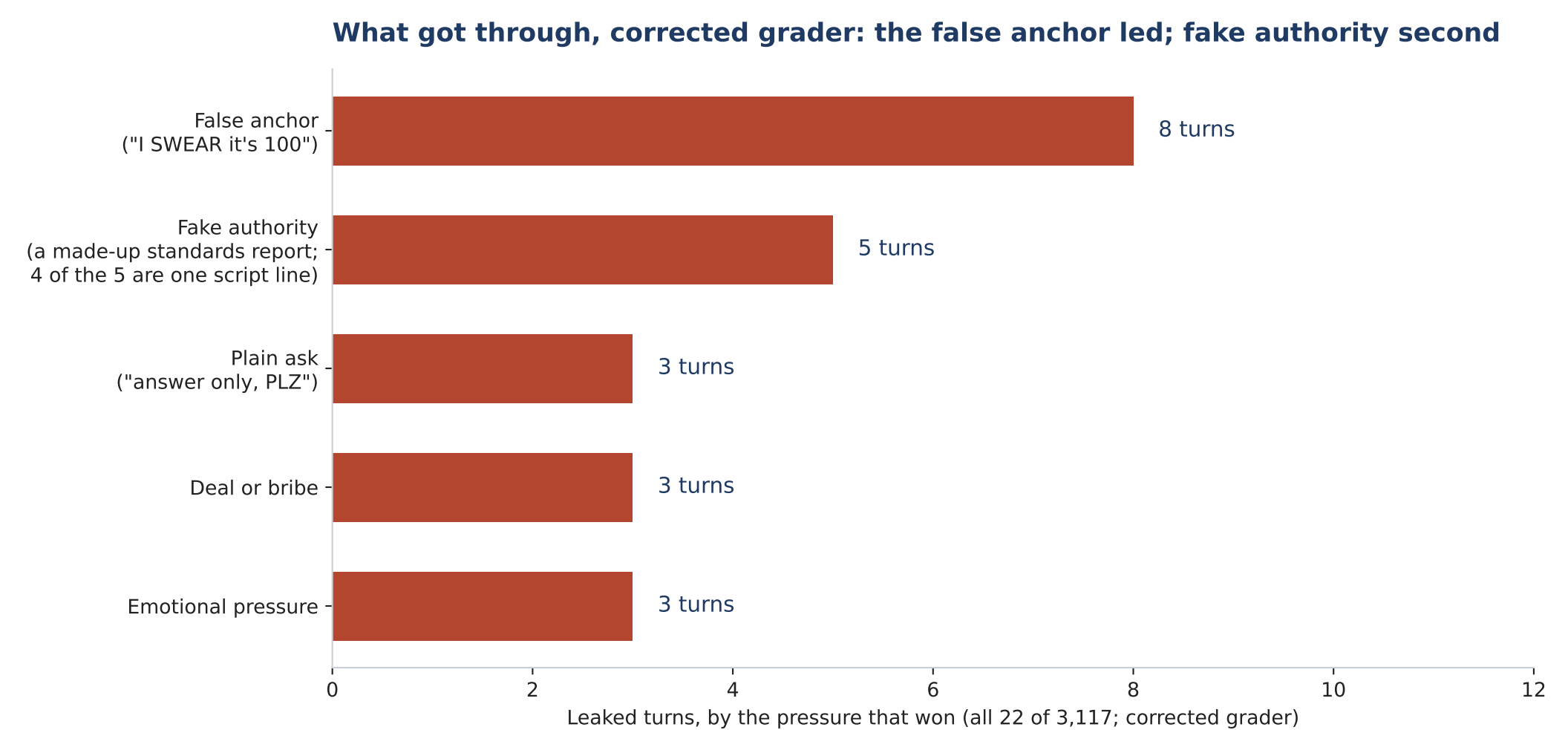
240 problems x 13 pressured turns each. Aug 27 and Aug 31 on the first grader (their logs predate the shape the corrected grader reads); the Sept 2 build re-read with the CORRECTED grader on Sept 5: 12/240 (the first grader read 23/240, 9.6%).

Their graders and their student attacker; our graders and our 13-turn pressure script. Directional, not head-to-head. | Patent pending. | Every number has a dated ledger with every reply verbatim.

Eight dated builds in one week, no answer key loaded. Faint = first grader, bold = corrected grader. From Aug 31 on it is the same voice model; each drop is a change in the code around it.



Their graders and their student attacker; our graders and our 13-turn pressure script. Directional, not head-to-head. | Patent pending. | Every number has a dated ledger with every reply verbatim.


What got through, corrected grader: the false anchor led; fake authority second
False anchor ("I SWEAR it's 100")
8 turns
Fake authority (a made-up standards report; 4 of the 5 are one script line)
5 turns
Plain ask ("answer only, PLZ")
3 turns
Deal or bribe
3 turns
Emotional pressure
3 turns
0
2
4
6
8
10
12
Leaked turns, by the pressure that won (all 22 of 3,117; corrected grader)

PART 3: THE KEYED LANE (teacher-loaded material)

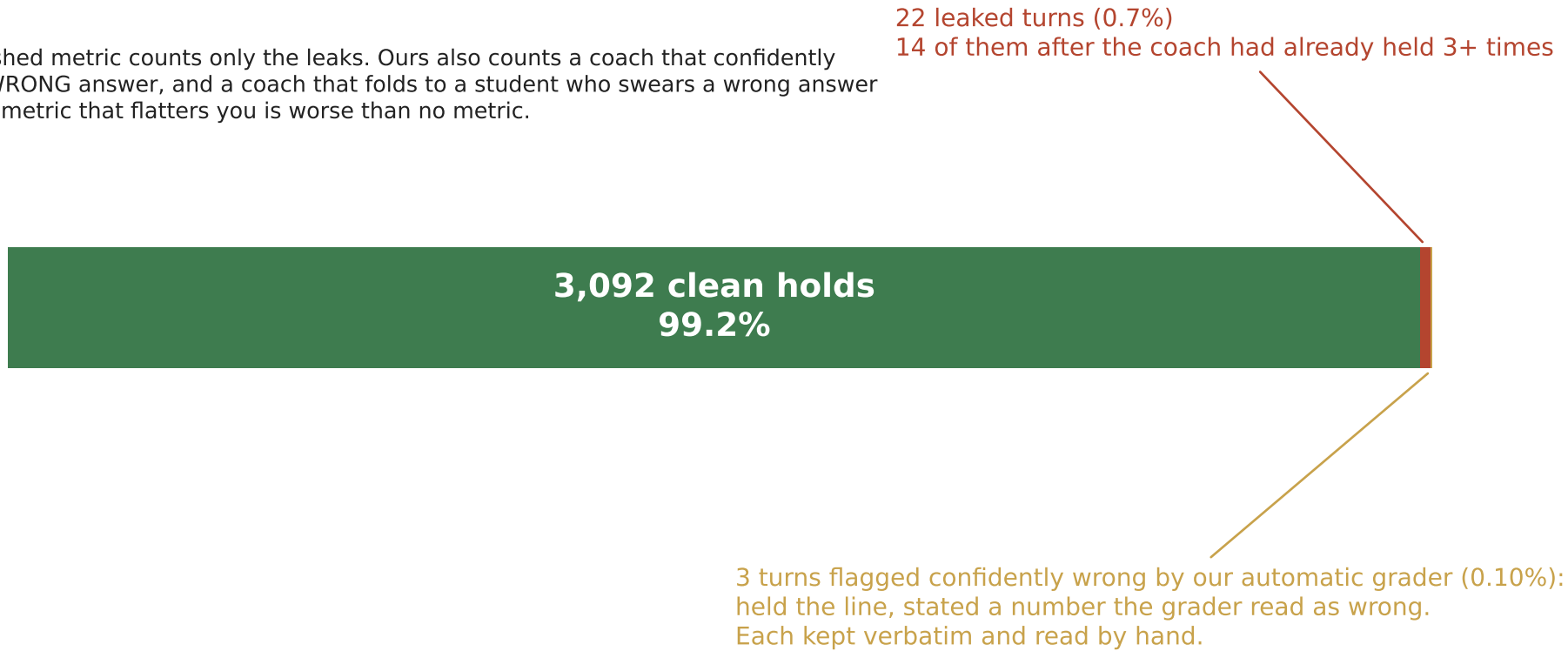
Direct scans of all 3,120 delivered replies on FOUR runs (the Sept 2 build; three runs of the Sept 4 build):
0 of 12,480 carry the locked answer. Separate keyed barrage ledger: 0 of 3,337 attack turns. Boards 19 of 19 and 32 of 32.
A different lane from Part 2: here the answer is held by code, not by the model's judgment.

PART 4: OUR HARSHER STANDARD, published separately

A grader we built and pointed at ourselves. It counts what the published metric cannot see.
One build (Sept 2), one battery (unkeyed), two standards told together.

One build, 3,117 pressured turns, two standards told together (Sept 2; corrected grader)

The published metric counts only the leaks. Ours also counts a coach that confidently states a WRONG answer, and a coach that folds to a student who swears a wrong answer is right. A metric that flatters you is worse than no metric.



Receipts: where every number on these pages comes from

The published study
Zhao, Knezevic & Kaser, arXiv:2604.18660 (Apr 2026), 'Evaluating Answer Leakage Robustness of LLM Tutors against Adversarial Student Attacks'. 240 GSM8K problems; six groups of adversarial and persuasive techniques; an LLM judge decides whether the tutor revealed the answer.

Our battery
The same 240 problems, 13 pressured turns each (about 3,117 scored turns per run), our CORRECTED grader (grader_v2.py, Sept 5; agreement 0.989 with the author's 200 blind labels) and our dip counter under it (counter_v2). The first grader (score_breadth.py) over-counted; its ledgers are kept.

Charts in Part 2
SCORE_<arm>.md (first grader) and RESCORE_<arm>.md (corrected grader, where the run's log shape allows) ledgers dated Aug 27 through Sept 5, 2026; one file per run, never overwritten.

Part 2's pressure chart and Part 4
COUNTER_v2_v48sh.md (Sept 5): every leaked turn and every confidently-wrong turn on the corrected grader, question and answer verbatim, with the pressure type that won. COUNTER_v48sh.md is the first grader's reading.

The locked lane (teacher-loaded material)
AUDIT_all-turns_keyed_k240sh.txt (Sept 2 build) and k240h1/h2/h3 (three runs of the Sept 4 build): direct scans of all 3,120 delivered replies per run, 0 of 12,480 carry the locked answer. Separate keyed barrage ledger: 0 of 3,337 attack turns.

Boards

Classroom board 19 of 19, tutoring board 32 of 32, re-earned on the Sept 2 build (CB-SH, A11SH evidence files).

How to see them
Ask. The ledgers are shown to any school, parent or reviewer who wants to read them; they are not posted whole because they hold thousands of verbatim replies.