Improving AI-Generated Officer Coaching for Community Corrections Supervision Visits

A one-iteration revision study: feedback-informed coaching updates, a cross-model test, and an exploratory benchmark against pooled expert human coaching

Ted Green and Madeline Warren (RePath Labs) — Independent expert reviewer and contributor: Valerie Meade

Version 1.0 · November 2026

Contents

Abstract....2
1. Executive summary....3
2. Scope and background....5
3. Methods....5
4. Results....8
5. Discussion....11
6. Limitations....13
7. Responsible use and recommendations for agencies....13
Data and ethics....13
Competing interests and funding....14
Author contributions and contributor statement....14
Acknowledgements....14
References....14
Appendix A: Review packet and evaluator-knowledge statement....15
Appendix B: Source identities (decoded post-return)....16
Appendix C: Supplementary results....16
Appendix D: Reproducing the numbers....18

Working paper (RePath Labs). Not peer-reviewed. This is a preprint made available for timely dissemination and comment. It has not undergone journal peer review; findings and framing may be revised in later versions.

Suggested citation: Green, T., & Warren, M. (2026). *Improving AI-Generated Officer Coaching for Community Corrections Supervision Visits: A one-iteration revision*

study with an exploratory benchmark against pooled expert human coaching. RePath Labs working paper, version 1.0, November 2026.

Companion papers:
Green, T., & Warren, M. (2026). *Validating AI-Generated Officer Coaching for Community Corrections Supervision Visits.* https://doi.org/10.31219/osf.io/m7d5g_v1
Green, T., Warren, M., & Meade, V. (2026a). *Establishing and Meeting a Reference Standard for AI Grading of Community Corrections Supervision Visits.* https://doi.org/10.31219/osf.io/xk5u9_v1
Green, T., Warren, M., & Meade, V. (2026b). *Validating an AI Grader for Community Corrections Supervision Visits.* https://doi.org/10.31219/osf.io/9wq32_v1

Correspondence: Ted Green, RePath Labs (tedgreen@ehawksolutions.com).

ORCID iDs: Ted Green 0009-0008-9264-6568; Madeline Warren 0009-0000-0141-8810; Valerie Meade (independent expert reviewer and contributor) 0009-0002-7261-0601.

License: © 2026 RePath Labs. All rights reserved.

Materials availability. The review rubric and its anchored score definitions, the decoded review scores, the analysis records, and model-version details are described in the appendices; the full verification package (frozen returned workbook, source key, decoded row-level table, analysis report, and verification records) may be made available to interested parties by agreement. The proprietary coaching generation prompts and the PII-bearing visit transcripts and human coaching comments are not distributed.

Scope. This tool is intended for officer coaching and outcome improvement. It is not intended for hiring, firing, promotion, discipline, or any other personnel decision.

Abstract

Validating AI-Generated Officer Coaching (Green & Warren, 2026) found that AI-generated officer coaching could stand with expert human coaching in blinded review: the strongest model's coaching was rated equal to or better than both directly compared expert human sources. That result opened an opportunity. AI coaching runs on a single method, so improving the method once improves the coaching on every scored visit an agency records. This paper reports one round of that improvement. We revised the coaching method using the expert reviewer's documented feedback and tested it on 24 different supervision visits (30 were selected; six were excluded before review for input quality). The same independent expert scored 120 notes across five conditions on the same eight criteria, with the

notes masked so she could not see who or what wrote each one. The revised method outscored the original on the 1–5 composite (3.906 vs. 3.760; higher on 15 visits, lower on nine). The number of notes rated ready to show an officer as written more than doubled, from 4 of 24 to 10 of 24. In an exploratory comparison requested after results, the revised method's mean rating stands 14.2% above the pooled mean of all 84 expert human coaching notes rated across the two studies (3.906 vs. 3.420). By this pooled measure, AI coaching now rates above the average of expert human coaching. The study also produced two revealing and unsurprising results. First, we applied the identical revised instructions to another very capable model, OpenAI's GPT-6 Astra. As we have repeatedly found, switching models on a specific task does not yield the same result: on the comparison the study was designed around, Astra's coaching scored 0.458 points below even the original Sonnet method. Validation attaches to a model-plus-method workflow, not to a prompt. Second, the expert human coaching in this study rated very high (4.130), reinforcing a pattern we have seen across all of our studies: expert humans show high capability and high variability. A better human coach is a real gain, but it is one improvement, limited to that coach and the visits they touch. An improvement to the AI coaching method carries to every scored visit an agency records. One iteration improved average coaching quality and the share of notes ready to show, and the improved method now rates 14.2% above the pooled human mean. That is the practical significance of this study.

1. Executive summary

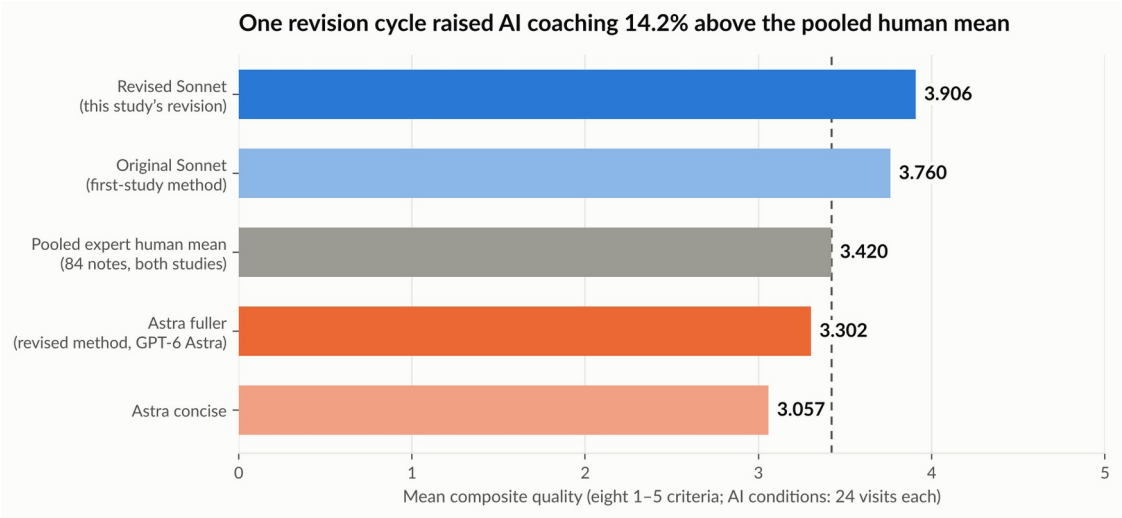


Figure 1. Mean composite quality (eight 1–5 dimensions) for this study's four AI conditions, each rated on 24 supervision visits by an independent expert who could not see which source wrote each note, alongside the pooled mean of all 84 expert human coaching notes rated across this study and its predecessor — an exploratory benchmark; full results and boundaries in §4.

Why we ran this study. *Validating AI-Generated Officer Coaching* (Green & Warren, 2026) established that AI-generated coaching could stand in blinded comparison with expert human coaching. It also handed us an opportunity. The blinded review documented exactly where the coaching fell short: mis-prioritization, unearned praise, missing concrete alternatives. Because the coaching runs on a single method, fixing those faults once would improve the coaching on every visit an agency records. This paper takes that opportunity: we revised the method from the expert's feedback and tested the revision, alongside the identical method on a second frontier model, on 24 visits none of the methods had seen.

What we found. The revision worked: the revised Sonnet method scored higher than the original on 15 of 24 visits (composite 3.906 vs. 3.760) and its notes rated ready to show as written went from 4 of 24 to 10 of 24. The gain is modest per note; its significance is structural. Improving one car's gas mileage saves a family a few dollars a week, while the same improvement across an entire fleet saves millions. A method improvement reaches every coaching note on every scored visit from that day forward, and nothing in this design suggests the cycle cannot be repeated. In an exploratory pooled benchmark, the revised method's mean now stands **14.2% above the mean of all 84 expert human coaching notes** rated across the two studies (3.906 vs. 3.420). The study also returned two revealing and unsurprising results. Switching models again produced a different result: the identical revised instructions on OpenAI's GPT-6 Astra scored 0.458 below even the original Sonnet workflow on the comparison the study was designed around, with no Astra note rated ready to show as written — models are not interchangeable parts. And the expert human coaching in this study rated very high (4.130, against 3.296 and 2.975 from the first study's human sources), reinforcing a pattern seen across all of our studies: expert humans show high capability and high variability.

The numbers worth remembering.

- **120 coaching notes, masked sources, one independent expert:** 24 visits × 5 conditions — original Sonnet (A), revised Sonnet (B), Astra fuller (C), Human Coach 1's authentic coaching (D), Astra concise (E). All 960 criterion ratings and 120 display judgments returned complete.
- **Composite quality:** Human Coach 1 **4.130**; revised Sonnet **3.906**; original Sonnet 3.760; Astra fuller 3.302; Astra concise 3.057.
- **One revision cycle, same visits:** the revised method scored higher than the original on 15 of 24 visits and lower on nine, an average gain of **+0.146** points. Notes rated ready to show as written went from **4 of 24 to 10 of 24**; notes rated unsuitable to show also rose from three to five (§4.2).
- **Pooled human benchmark (exploratory, post hoc):** 84 notes, mean 3.420; revised Sonnet 3.906; **+0.487, or 14.2%**, computed before rounding.

- **Primary comparison (set before review):** Astra fuller minus Original Sonnet **−0.458**; the revised method on a different model did not improve on the original workflow. Every sensitivity preserves every comparison's direction (§4.5).

What this is for. Officer coaching: short, transcript-grounded, standards-linked guidance generated for every scored visit and delivered through a filtered cadence. The display filter remains essential: 14 of 24 revised-Sonnet notes still needed editing or suppression. **What this is not for:** hiring, firing, promotion, discipline, or any other personnel decision. The validation does not authorize those uses.

2. Scope and background

The coaching tool is a practice-improvement layer on top of visit-quality measurement: it reviews the transcript and score context, evaluates the visit for coaching, and provides a short officer-facing note. Its intended and out-of-scope uses are unchanged from *Validating AI-Generated Officer Coaching* (Green & Warren, 2026), which carries the full scope statement and the supporting literature. In brief: the evidence chain links officer fidelity to core correctional practices with lower recidivism (Labrecque et al., 2023; Chadwick et al., 2015; Bonta et al., 2021); *Validating an AI Grader* validated measuring visit quality at scale (Green, Warren, & Meade, 2026b); and *Validating AI-Generated Officer Coaching* validated turning those measurements into expert-acceptable coaching.

This study tests the link that makes the chain durable: whether the coaching can be systematically improved. The case for caring about a modest gain is arithmetic. Officer training works and its effect decays, so sustaining one human coach requires recurring investment per coach. An AI coaching method is a single artifact: improve it once and the improvement applies to every note it generates — every officer, every visit, every agency running the workflow — immediately and without decay. The first study's blinded review supplied the targets (mis-prioritization, unearned praise, missing concrete alternatives), and the revision tested here was built directly from them.

This study also sharpens a scope boundary: the validation is workflow-specific. Running the same instructions on a different model produced worse coaching than the original method, so no result here extends to any untested model-method combination.

3. Methods

3.1 Sample and common inputs

Thirty visits were selected from the historical corpus, separate from the 30 reviewed in the first coaching study; six documented input-quality exclusions left

24 eligible visits: ten adult supervision, six juvenile, and eight treatment court. Historical AI-score strata informed selection. No visit IDs overlap the first study's 30, but three visits share anonymized case IDs with visits used during method development, and the 24 visits map to 22 case IDs (two repeated-case pairs); case IDs are not verified unique people, and officer linkage is unknown. This is a new-visit test of the methods, not an external holdout; §4.5 reports the overlap sensitivities. All AI conditions received identical inputs per visit: the transcript, the skill definitions, and common grading context generated with the accepted procedure from *Establishing and Meeting a Reference Standard* (Green, Warren, & Meade, 2026a). The original coaching method ran on these updated common inputs, so condition A is the original *method* on current inputs, not an exact replay of the first study's outputs. Human notes and ratings from the final cohort were excluded from all coaching development and generation inputs.

3.2 Five conditions

A — Original Sonnet. The first study's executed coaching method (claude-sonnet-4-6, 2,500-token cap), unchanged, on the common inputs: the baseline the revision must beat.

B — Revised Sonnet. The feedback-informed one-step v3.2 method on the same model (claude-sonnet-4-6, temperature 0.2, extended thinking disabled, 4,096-token cap), targeting the first study's documented failure modes: select the most consequential behavior-change priority, ground it in the transcript, give a concrete alternative action, and avoid praise that overstates the visit.

C — Astra fuller. The identical v3.2 method and inputs on OpenAI's GPT-6 Astra (isolated gpt-6-astra workers at maximum reasoning effort, no tools). This is the cross-model test and the pre-specified primary contrast (C–A): if methods were portable, a newer model running the better method should beat the old workflow.

D — Human Coach 1. Maureen Flanagan's complete authentic general coaching comments for these visits, written during her ordinary expert review of the new corpus; text not rewritten; de-identified as HC1, consistent with the first study. She completed that review naive to the coaching comparison: she was not aware of the first study's coaching evaluation data, had not read the published study, and had not been informed of its results before or during her review; she confirmed this in writing upon returning her packet, and the published paper was provided to her only afterward.

E — Astra concise. A shorter presentation generated jointly with C under instructions to preserve the same core advice and qualifications; E–C evaluates the implemented presentation strategy, not an isolated effect of word count.

The revision was drafted against the first study's archived ratings and comments with formative internal review, then frozen as the one-step v3.2 method and evaluated with its known weaknesses rather than tuned against the final reviewer results. Model evaluators tested against archived ratings did not qualify for unattended method selection; an evidence-plan-then-writer pilot was not adopted; unsuccessful pilots and original outputs remain in the study record.

3.3 Execution and expert review

There were 144 authorized first calls (72 grading, 48 Sonnet coaching, 24 Astra pairs). Two Astra calls failed the declared output-format rule, affecting four note slots on two visits; one authorized exact-input retry per visit filled them before review (146 total calls), and §4.5 reports the first-attempt sensitivity. No successful coaching text was replaced or manually improved.

The packet was sent September 22 and returned October 25, 2026. Valerie Meade reviewed all 120 notes with rotating Coach A–E labels masking sources within each visit; prior scores and the source key were withheld. She rated the eight unchanged anchored 1–5 criteria from the first study and made a four-category display decision per note, with no grading curve or forced ranking; visit-level preferences were optional and permitted ties and "none suitable." All 960 criterion ratings and 120 display judgments were returned complete. **Evaluator knowledge:** this round is an informed-evaluator design, disclosed rather than obscured — the evaluator had access to the decoded results of the prior study; source labels were withheld, but source-type inference was plausible. She also flagged nearly identical wording between two notes on two visits, in both cases the jointly generated Astra pair (Appendix A; §6). **Bridging repeats:** eight unchanged historical notes from the first study, selected before final-output inspection, were re-rated with the same criteria plus a recognition question, prior scores withheld; they are reported in §4.5 as context for cross-round comparability and do not enter the primary dataset.

3.4 Analysis plan and registration status

Each composite is the unweighted mean of the eight ratings; the visit is the unit for paired comparisons. A written analysis plan dated September 22, 2026 — finalized before the reviewer's packet was returned — fixed the four contrasts (C–A primary; B–A, C–D, E–C supporting), the two separate sensitivities, the first-attempt reporting rule, and a descriptive approach: no significance tests, population intervals, imputation, or recalibration. The study was not publicly preregistered; the September 22 plan, preserved unchanged in the study record, is its pre-specification. The pooled human benchmark and the Sonnet-versus-HC1 exploration were requested after results were available and are labeled exploratory throughout.

4. Results

4.1 All five conditions

Table 1. Condition scores and display decisions (24 fully rated notes per condition)

Condition	Mean	Median	Range	Y / E / D / N
A Original Sonnet	3.760	3.875	2.625–4.750	4 / 17 / 0 / 3
B Revised Sonnet	3.906	4.063	2.250–5.000	10 / 9 / 0 / 5
C Astra fuller	3.302	3.313	2.250–4.625	0 / 12 / 0 / 12
D Human Coach 1	4.130	4.250	3.125–4.875	12 / 12 / 0 / 0
E Astra concise	3.057	3.000	2.250–4.625	0 / 11 / 0 / 13

A–E are analysis labels, not the workbook's rotating reviewer labels. Y = ready as written; E = edit first; D = digest only; N = do not show next to score. Ready-as-written proportions: A 16.7%, B 41.7%, C 0%, D 50.0%, E 0%. "Edit first" is not ready. In plain language: the revised Sonnet method is the best AI condition and the only one whose readiness rate approaches the human comparator's; Human Coach 1's authentic coaching leads the study and never drew a do-not-show judgment; neither Astra variant produced a single note the expert would display as written.

4.2 The planned within-visit contrasts

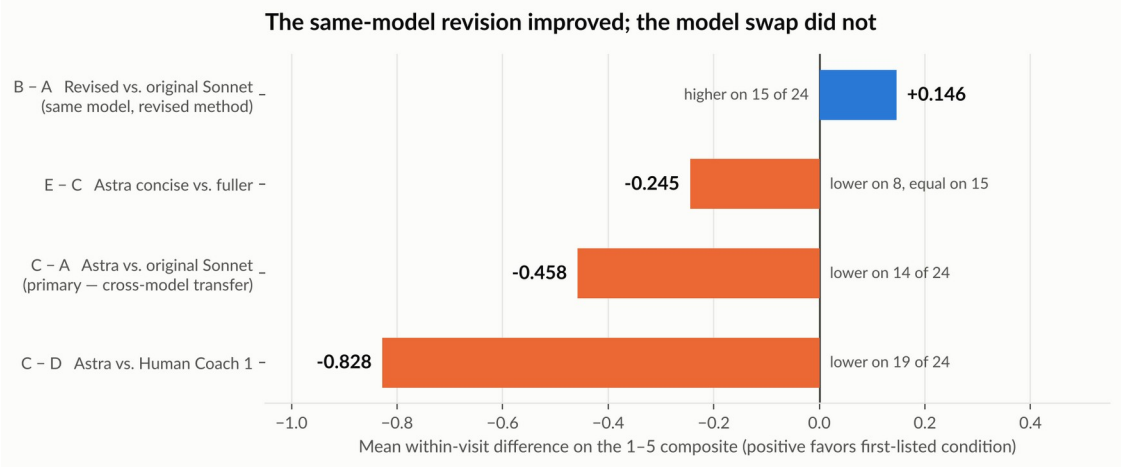


Figure 2. The four pre-specified within-visit contrasts (mean paired difference on the 1–5 composite; positive favors the first-listed condition). The one planned improvement that materialized is the same-model revision (B–A). The primary contrast, the revised method on a different model (C–A), did not improve on the original workflow.

Table 2. Planned within-visit differences (24 complete pairs each)

Contrast	Mean	Median	Range	Higher / equal / lower
Astra fuller – Original Sonnet (primary)	−0.458	−0.438	−1.875 to +1.000	9 / 1 / 14
Revised Sonnet – Original Sonnet	+0.146	+0.313	−1.750 to +1.625	15 / 0 / 9
Astra fuller – Human Coach 1	−0.828	−0.938	−2.250 to +0.750	4 / 1 / 19
Astra concise – Astra fuller	−0.245	0.000	−1.625 to +0.125	1 / 15 / 8

The revision worked (B–A). The revised method beat the original on 15 of 24 visits with a mean gain of +0.146, and its ready-as-written count went from 4 to 10. The gain was uneven: the do-not-show count also rose from three to five, so the revision raised the quality of its best notes more reliably than it protected its worst. Where it missed, the reviewer's comments describe the same residual failure modes the revision targeted, now less frequent but not eliminated.

Switching models produced a different result (C–A, primary). The identical revised method on GPT-6 Astra scored below the *original* Sonnet workflow on 14 of 24 visits, mean −0.458; the pre-specified improvement proposition was not supported. Against the human comparator the gap widened to −0.828 (C–D, lower on 19 of 24).

Concise was not free (E–C). The concise Astra variant rated identical to its fuller sibling on 15 visits, lower on eight, higher on one (mean −0.245). A zero median does not establish equivalence; what the data rule out is the hope that compression came at no cost.

4.3 The criterion profile

Table 3. Mean rating for every criterion (24 ratings per cell)

Criterion	A	B	C	D	E
Standards alignment	2.750	3.125	2.042	4.000	1.833
Transcript grounding	3.292	3.917	3.417	3.583	3.375
Prioritization	2.917	3.167	2.208	4.250	1.958
Usefulness/actionability	3.292	3.375	2.875	3.125	2.333
Tone/receptivity	4.792	4.708	4.333	4.417	3.875
Brevity/cognitive load	4.958	4.875	4.750	4.917	4.750
Safety/overclaiming	4.583	4.458	4.208	4.583	4.000
Subsection/display fit	3.500	3.625	2.583	4.167	2.333

Three readings matter. The revision's largest gain over the original was transcript grounding (+0.625), with standards alignment (+0.375) and prioritization (+0.250) also up; actionability barely moved (+0.083), the clearest target for the next

iteration. Astra's deficit is substantive, not stylistic: tone, brevity, and safety are competitive while standards alignment, prioritization, and display fit are far below every comparator — the reviewer's comments describe coaching aimed at trivial or compliance topics while consequential behavior-change material went unaddressed. And the human comparator leads on standards alignment and prioritization but not on actionability, where revised Sonnet's mean is the highest in the study — the expert repeatedly asked the human notes for more specific alternatives, the exact dimension the revision was built to strengthen.

4.4 Exploratory benchmark: revised Sonnet vs. all rated human coaching

Table 4. Pooled expert human coaching across both studies (exploratory, post hoc)

Source and review sample	Notes	Mean / 5
Human Coach 2, first study	30	2.975
Human Coach 1, first study	30	3.296
Human Coach 1, current study	24	4.130
Pooled human benchmark	**84**	**3.420**
Revised Sonnet, current study	**24**	**3.906**

Giving each human note equal weight, the pooled mean is 3.419643. Revised Sonnet's mean is 3.906250, a difference of +0.486607, or **14.2%** above the pooled human mean, computed before rounding (100 × (3.906250 − 3.419643) / 3.419643 = 14.23).

This is the research program's question in one number, and it carries both weight and boundaries. The weight: across 84 expert human coaching notes spanning two studies, three review occasions, and 54 distinct visits, the average expert human note rates 3.42, and after one revision cycle the AI method's notes rate 14.2% above that — produced on demand, for every visit, at marginal cost near zero. We benchmark against the pooled mean because it represents the expert human coaching actually observed across the series — what an agency's officers collectively receive — rather than a single coach's best occasion. The boundaries: the comparison was requested after results were available; the pool contains two human sources (HC1 contributes 54 notes, HC2 30), not a population of coaches; the 84 notes are not independent observations; the two studies' visit samples were drawn from the same corpus under similar score-informed selection but were not matched on difficulty; and a 1–5 scale has no meaningful zero, so 14.2% describes the numerical means — not "14.2% better coaching," a standardized effect size, or any outcome claim. Same-corpus anchors qualify it honestly: against the first study's original Sonnet mean (3.575) the revised method is +0.331, though its ready-as-written rate (10/24, 41.7%) is not higher than that earlier round's (13/30, 43.3%);

and within the current study, revised Sonnet remains 0.224 below Human Coach 1, exceeding her composite on nine visits, tying on two, and trailing on 13, while averaging higher on grounding, actionability, and tone.

4.5 Sensitivities, bridging repeats, preferences, and cost in brief

Full tables for this section are in Appendix C. **Sensitivities:** excluding the two Astra retry visits, the three development-case-overlap visits, or the one intake-flagged visit — each separately, per the analysis plan — preserves the direction of every specified contrast; B–A strengthens to +0.228 when the intake-flagged visit is excluded. **Bridging repeats:** on the eight unchanged historical notes, re-rated with prior scores withheld, the mean rose from 3.313 to 3.516 (+0.203; five rose, three fell; mean absolute change 0.391); criterion agreement was within one point on 58/64 cells (90.6%); five of eight notes were recognized; and display judgments moved only toward strictness (three Yes became Edit first, one Edit first became do-not-show). Scores this round were not systematically harsher while the display bar was at least as strict, which argues against attributing revised Sonnet's readiness gain to a softer standard. **Preferences (optional, as entered):** the ordering matches the composites — Human Coach 1 selected on 14 visits, revised Sonnet on eight (five sole selections, against the original method's zero), with four "none suitable" visits; four entries carry interpretive tensions that are preserved rather than recoded, and preference counts would warrant neutral clarification before serving as primary evidence. **Length:** revised Sonnet's notes are longer than the original's (179.5 vs. 113.8 words on average), worth watching as readiness pressure.

5. Discussion

One iteration worked, and the arithmetic is the point. A single feedback-informed revision cycle measurably improved AI coaching on visits the method had never seen, with the largest gains on exactly the dimensions the revision targeted. Combined with the exploratory pooled benchmark, the study moves the program's central claim forward: AI coaching at every-visit scale now rates at or above the expert human average, and the improvement loop that got it there is repeatable. The right way to price the result is not "+0.146 points" but "+0.146 points on every coaching note the system will ever generate, for one review-and-revise cycle." The same arithmetic sets the next cycle's targets: actionability barely moved, and the do-not-show floor did not improve.

Models are not interchangeable parts. The pre-specified primary contrast came back negative, a revealing and unsurprising result: it repeats, in the coaching layer, what this project has found every time it has swapped models under a fixed task. The identical instructions, transcripts, and grading context that produced the study's best AI coaching on Sonnet produced coaching on Astra that the expert

would not display as written even once in 24 visits, with the deficit concentrated in substance (prioritization, standards alignment, display fit), not style. For agencies and for RePath's own engineering the implication is the same: a validated workflow is a model-plus-method unit, and every model swap, however routine it looks, re-opens validation.

Expert humans: high capability, high variability. Human Coach 1's current coaching (4.130) outscored her first-study notes (3.296) and led this study; across the series, expert human source-occasions span 2.975 to 4.130. This is the picture our studies keep drawing of expert human performance, in grading as in coaching: capable of the best work in the corpus, and widely variable across people and occasions. The current rise has two candidate explanations — a different review context, and genuine growth from her calibration work in *Establishing and Meeting a Reference Standard* (Green, Warren, & Meade, 2026a) — neither established. The bridging repeats bound a third explanation rather than eliminate it: on identical historical notes the reviewer scored +0.203 higher this round, so some general drift is present — but it is roughly a quarter of Human Coach 1's +0.834 rise, and the reviewer's display judgments moved only toward strictness. The variability is the program's point. The thesis has never been that AI coaching should beat the best human experts; it is that validated AI coaching can give every officer, after every visit, feedback at or above the expert human average — consistently, where human coaching quality depends on which coach and which occasion, and where in-person feedback often goes unwritten, unremembered, and untracked. The scalability asymmetry is the practical one: the AI method produced each note in about half a minute of recorded compute and can do so for every visit an agency records, while expert human coaching is bounded by expert hours and calendars — which is why most supervision visits today receive no written coaching at all.

The display filter stays. Revised Sonnet's 10/24 ready-as-written rate is the best AI rate observed in the program and still leaves 14 of 24 notes that invite human review before display. The human comparator's floor is also worth naming: her coaching never drew a do-not-show judgment, while the revised method drew five — the clearest remaining quality gap, and the reason the filter sits between generation and officers. Unattended display is not established and is not proposed. The supported configuration is unchanged from *Validating AI-Generated Officer Coaching* (Green & Warren, 2026): generate for every scored visit, display as-written only where the note clears the bar, route the rest through edit or digest cadences. The concise-presentation result cautions against compressing notes without re-validation.

6. Limitations

One expert reviewer rated all notes; her judgments are the study's evidence, and a second independent reviewer would strengthen any future claim. This round was an informed-evaluator design: the evaluator had access to the decoded results of the prior study; source labels were withheld, but source-type inference was plausible, and she recognized the jointly generated Astra pairs on two visits by their near-identical wording — masking did not guarantee independent evaluation of every note. Five of eight historical repeats were recognized. The sample is 24 non-randomly selected visits from the same historical corpus, with three development-case overlaps, two repeated-case pairs, unverified person identity, and unknown officer linkage; it is not an external holdout, and the sensitivities (§4.5, Appendix C) mitigate but cannot eliminate these concerns. The analysis is descriptive under a written internal pre-review plan rather than a public preregistration. The pooled human benchmark is exploratory and post hoc, pools non-independent notes from two sources across differing review occasions, and reads a ratio from a scale with no meaningful zero. The equal-weighted composite is a practical index, not a validated psychometric scale. Model, prompt, grading-context, and presentation effects are not isolated; the original-method arm ran on updated common inputs rather than replaying the first study. Nothing here measures officer behavior change, supervision decisions, or participant outcomes, and neither unattended display readiness nor superiority to typical field coaching is established.

7. Responsible use and recommendations for agencies

The recommendations of *Validating AI-Generated Officer Coaching* carry forward unchanged, with one addition. Unchanged: use the coaching layer for officer development, not personnel decisions; keep the display filter between generation and officers; keep humans in the loop for editing and digests; and treat expert-rated quality as necessary but not sufficient evidence pending field outcomes. The addition this study compels: **treat the model as part of the validated configuration.** An agency or vendor that changes the underlying model — even to a newer, more capable one — should treat the coaching layer as unvalidated until it has been re-tested against expert review on held-out visits. This study's primary result is a direct demonstration of why.

Data and ethics

The visit transcripts are de-identified recordings of real supervision visits used for this research under RePath's existing agency contracts, with participant and officer identifiers removed before any grading or coaching generation, drawn from multiple community corrections agencies. The 24-visit cohort is distinct from the first coaching study's 30 visits. The human coaching comments evaluated here

were written by Human Coach 1 during her ordinary expert review of the new corpus; she consented to the use of her comments and to being thanked by name. The external expert reviewer was compensated for her time, reviewed only de-identified text with source labels masked, and consented to be named. No interventions or operational decisions were assigned in this study.

Competing interests and funding

RePath Labs designed, funded, and conducted this study and has a commercial interest in officer-visit grading and coaching tools. There was no external funding. Valerie Meade served as the independent expert reviewer in a compensated consulting role; her review was conducted with source labels masked, under the evaluator-knowledge disclosure in §3.3 and Appendix A.

Author contributions and contributor statement

Ted Green led project framing, study design, model-output generation, analysis, and drafting. **Madeline Warren** contributed to study design, data organization, reviewer-packet preparation, source-material coordination, and paper development. **Valerie Meade** (contributor and independent expert reviewer) performed the expert review of all 120 coaching notes and the eight historical repeats. She is credited as a contributor rather than an author deliberately: because her ratings are the study's evidence, keeping the evaluation role separate from authorship of the conclusions preserves the independence of the review. Consistent with that separation, her review of this manuscript was limited to the factual accuracy of the descriptions of her credentials and review process; she had no role in the analysis, interpretation, or drafting of conclusions. Expert grader **Maureen Flanagan** is acknowledged as the human coaching source; her comments were written during her independent review of the new corpus and used here with the study's disclosures. AI tools assisted development, execution, analysis, and drafting.

Acknowledgements

The authors thank Valerie Meade for performing the expert review at the center of this study, and Maureen Flanagan, whose authentic coaching comments served as the human comparison source and the study's quality benchmark. Katie Myers's coaching comments from the first study contribute to the pooled human benchmark. All three experts' sustained, careful work underlies this series.

References

Bonta, J., Bourgon, G., Rugge, T., Pedneault, C. I., & Lee, S. C. (2021). A system-wide implementation and evaluation of the Strategic Training Initiative in Community

Supervision (STICS). *Journal of Criminal Justice*, 74, 101816. https://doi.org/10.1016/j.jcrimjus.2021.101816

Chadwick, N., Dewolf, A., & Serin, R. (2015). Effectively training community supervision officers: A meta-analytic review of the impact on offender outcome. *Criminal Justice and Behavior*, 42(10), 977–989. https://doi.org/10.1177/0093854815595661

Green, T., & Warren, M. (2026). *Validating AI-Generated Officer Coaching for Community Corrections Supervision Visits: A blinded expert-review study of standards-grounded coaching generated from visit transcripts and scores.* RePath Labs working paper, version 1.0.1, September 2026. https://doi.org/10.31219/osf.io/m7d5g_v1

Green, T., Warren, M., & Meade, V. (2026a). *Establishing and Meeting a Reference Standard for AI Grading of Community Corrections Supervision Visits: An expert adjudication and multi-model agreement study.* RePath Labs working paper, version 1.0, September 30, 2026. https://doi.org/10.31219/osf.io/xk5u9_v1

Green, T., Warren, M., & Meade, V. (2026b). *Validating an AI Grader for Community Corrections Supervision Visits: A reliability and agreement study against expert human auditors.* RePath Labs working paper, version 1.0, July 2026. https://doi.org/10.31219/osf.io/9wq32_v1

Labrecque, R. M., Viglione, J., & Caudy, M. (2023). The impact of community supervision officer training programs on officer and client outcomes: A systematic review and meta-analysis. *Justice Quarterly*, 40(4), 587–611. https://doi.org/10.1080/07418825.2022.2120062

RePath Labs. (2026). *Follow-up study analysis plan and completed-set generation records (September 22, 2026); Analysis report and coaching diagnosis (October 26, 2026).* Restricted study record.

Appendix A: Review packet and evaluator-knowledge statement

The reviewer packet (sent September 22, 2026; returned October 25, 2026) contained: 24 visit sections, each with the de-identified transcript reference, common grading context, and five coaching notes under rotating Coach A–E labels; a separate historical-repeats section with eight unchanged first-study notes and a recognition question per note; the eight anchored 1–5 criterion definitions, unchanged from the first study; the four display categories with definitions; an optional per-visit preference item permitting ties and “none suitable”; and an unchanged skills-reference appendix. Prior scores, source identities, and the source key were withheld.

This round is an informed-evaluator design, and the paper reports it as such: the evaluator had access to the decoded results of the prior study; source labels were withheld, but source-type inference was plausible. Within-visit Coach A–E labels rotated so that no label tracked a source across visits. The reviewer flagged nearly identical wording between two coaches on two visits; both were the jointly generated Astra fuller/concise pairs, so masking did not guarantee independent evaluation of paired notes. The human coaching source (Human Coach 1) completed her independent review of the new corpus naive to the first study's coaching-comparison results, confirmed in writing upon returning her packet; the published first study was provided to her only afterward.

Appendix B: Source identities (decoded post-return)

Analysis labels decode as: **A** — original-method coaching on Anthropic claude-sonnet-4-6 (the first study's executed method, 2,500-token cap, on current common inputs); **B** — revised one-step v3.2 method on claude-sonnet-4-6 (temperature 0.2, extended thinking disabled, 4,096-token cap); **C** — one-step v3.2 fuller method on OpenAI gpt-6-astra (isolated workers, maximum reasoning effort, no tools); **D** — Human Coach 1 (Maureen Flanagan), complete authentic general coaching comments, text not rewritten; **E** — one-step v3.2 concise presentation on gpt-6-astra, jointly generated with C. Sonnet's auxiliary concise outputs were generated but were not a reviewed condition. The original Sonnet outputs' internal model-name strings disagreed with the provider label; preserved execution records identify the provider model, and self-reported strings do not establish a different model.

Appendix C: Supplementary results

Table C1. Separate sensitivity comparisons (exclusions never combined)

Subset	n	Pair	Mean	Median	Range	+ / = / −
First attempts	22	C–A	−0.449	−0.438	−1.875 to +1.000	8 / 1 / 13
First attempts	22	C–D	−0.841	−0.938	−2.250 to +0.750	3 / 1 / 18
First attempts	22	E–C	−0.199	0.000	−1.250 to 0.000	0 / 15 / 7
No dev. overlap	21	C–A	−0.429	−0.500	−1.750 to +1.000	8 / 1 / 12
No dev. overlap	21	B–A	+0.131	+0.250	−1.750 to +1.625	13 / 0 / 8
No dev. overlap	21	C–D	−0.875	−1.000	−2.250 to +0.750	4 / 0 / 17
No dev. overlap	21	E–C	−0.232	0.000	−1.625 to +0.125	1 / 14 / 6
No intake flag	23	C–A	−0.402	−0.375	−1.875 to +1.000	9 / 1 / 13
No intake flag	23	B–A	+0.228	+0.375	−1.125 to +1.625	15 / 0 / 8
No intake flag	23	C–D	−0.777	−0.875	−2.250 to +0.750	4 / 1 / 18
No intake flag	23	E–C	−0.255	0.000	−1.625 to +0.125	1 / 14 / 8

First attempts excludes the two Astra retry visits (B–A was not specified for that subset); no dev. overlap excludes the three development-case-overlap visits; no intake flag excludes the one visit with an administrative intake discrepancy.

Table C2. Re-rated historical notes from the first study (prior scores withheld)

Note	Earlier	Current	Change	Display transition	Recognized
H01	2.375	3.375	+1.000	No to No	Yes
H02	3.875	3.750	−0.125	Edit to Edit	Yes
H03	3.250	3.625	+0.375	Yes to Edit	Yes
H04	3.000	3.125	+0.125	No to No	Yes
H05	3.000	2.750	−0.250	Edit to No	Yes
H06	4.125	3.750	−0.375	Yes to Edit	No
H07	3.625	3.875	+0.250	Yes to Edit	No
H08	3.250	3.875	+0.625	Edit to Edit	No

“No” abbreviates “Do not show next to score.” Criterion agreement: exact on 27/64 cells, within one point on 58/64. Eight selected repeats cannot establish reviewer reliability or explain why individual judgments changed; no scores were recalibrated.

Table C3. Visit-level preference selections (optional; ties permitted)

Condition	Selected at all	Selected alone	With another
A Original Sonnet	3	0	3
B Revised Sonnet	8	5	3
C Astra fuller	2	1	1
D Human Coach 1	14	7	7
E Astra concise	1	0	1

Thirteen visits had a single selection, seven multiple, four “none suitable as written”; none were blank or marked “no meaningful difference.” Four entries carry interpretive tension (a comment favoring one of three tied selections; a sole selection of a do-not-show note; a none-suitable choice despite one Yes-rated note; a tie between notes with sharply different scores); all are preserved as entered.

Table C4. Note length and recorded first-call latency

Condition	Words mean	Median	Word range	Latency median (s)
A Original Sonnet	113.8	106.5	86–177	21.626
B Revised Sonnet	179.5	180.5	148–217	28.811
C Astra fuller	149.4	149.5	140–160	219.567
D Human Coach 1	98.0	95.0	56–174	Not measured
E Astra concise	92.4	93.0	81–100	Shared C call

Astra retries took 241.504 and 309.732 seconds. Routes and call scope differ; latencies are not a controlled model-speed comparison.

Appendix D: Reproducing the numbers

All statistics in this paper recompute from the frozen study record. Each note's composite is the unweighted mean of its eight 1–5 ratings; paired contrasts difference composites within visit and classify before rounding; display counts tally the four categories as entered. The pooled human benchmark is (30 × 2.975 + 30 × 3.296 + 24 × 4.130) / 84 = 3.419643 against revised Sonnet's 3.906250: +0.486607, 14.23%. The returned workbook's SHA-256 fingerprint is published so that any later recipient of the restricted verification package can confirm the file behind this paper's numbers is unchanged: c185c406032d8c9015dada81b145b03c43b0a4a5e90b160286a6b08d2c20194c; results.json preserves full-precision values and input hashes; verify_analysis.py independently recomputes every reported statistic, checks all 120 source assignments against preserved generation and human-source records, and checks all 64 historical prior scores against the original first-study workbook snapshot. The September 22, 2026 analysis plan is retained unchanged in the study record. This paper's tables were additionally re-verified against the raw decoded return on October 27, 2026, and again at version 1.0 on November 14, 2026.