

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

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

Version 1.0.1 · September 2026

Contents

Abstract.....	2
1. Executive summary.....	3
2. Scope and intended use.....	6
3. Background: from visit measurement to officer coaching.....	7
4. Methods.....	8
5. Results.....	11
6. Discussion.....	17
7. Limitations.....	18
8. Future work.....	19
9. Responsible use and recommendations for agencies.....	20
Data and ethics.....	20
Competing interests and funding.....	20
Author contributions and contributor statement.....	21
Acknowledgements.....	21
References.....	21
Appendix A: Review packet inventory.....	22
Appendix B: Blinding statement (as given to the reviewer).....	23
Appendix C: Source identities (decoded post-return).....	23
Appendix D: Reproducing the numbers.....	23

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 Not peer-reviewed.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). *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. DOI: 10.31219/osf.io/m7d5g_v1.

Companion paper: Green, T., Warren, M., & Meade, V. (2026). *Validating an AI Grader for Community Corrections Supervision Visits: A reliability and agreement study against expert human auditors*. RePath Labs working paper. DOI: 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 blinded-review scores, the analysis script, and model-version details are described in Appendix A and Appendix D; the full verification package (frozen returned workbook, source key, decoded row-level table, analysis report, and claims ledger) may be made available to interested parties by agreement. The proprietary coaching generation prompt 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

Meta-analytic evidence links sustained officer fidelity to core correctional practices with substantially lower participant recidivism; caseloads supervised by untrained officers recidivate roughly 38% more often than those supervised by trained officers (Chadwick et al., 2015), and fidelity after training — not the training event itself — is the moderator that makes the difference (Labrecque et al., 2023). Our companion study validated the measurement layer: a frontier large language model can grade recorded supervision visits against a standards-based rubric about as reliably as expert human auditors (Green, Warren, & Meade, 2026). This paper tests the practice-change layer: whether an LLM can turn those transcripts and scores into officer-facing coaching that an independent expert judges to be as good as coaching written by expert human reviewers. On the same 30-visit corpus, we

assembled 120 coaching outputs — four per visit, from two expert human sources (authentic prior grader coaching notes) and two LLM sources (Claude Sonnet 4.6 and GPT-5.5), randomized within visit — and had a third expert, Valerie Meade, score every output on eight pre-specified 1–5 dimensions, rank the four outputs within each visit, and make a product-display decision. The review was blinded: source identities and source types were withheld until scoring was complete, and the reviewer’s own historical coaching was excluded from the packet. Decoded against the internal source key, the strongest LLM source (Claude Sonnet 4.6) was rated equal to or better than both directly compared human sources on the composite quality index (vs. one human source +0.60, Holm-adjusted $p = 0.006$; vs. the other +0.28, not significantly different and formally non-inferior at a 0.4-point margin, one-sided $p = 0.0002$), earned the best mean within-visit rank (2.07 of 4), the most first-place outputs (10 of 30), and the most display-ready outputs (13 of 30). Pooled by type, LLM coaching outscored human coaching on the visit-level composite (3.44 vs. 3.14; paired difference +0.30, 95% CI +0.10 to +0.52, $p = 0.019$). We report the result with its boundaries: one expert reviewer, two directly compared human sources, non-standardized human comparison text, a 30-visit corpus, and no claim of field behavior change or recidivism impact. Together with the grader validation, the finding supports the working thesis that agencies can responsibly act on validated visit scores with display-filtered LLM-generated coaching — the mechanism the training literature ties to lower recidivism — while the outcome link itself awaits field study.

1. Executive summary

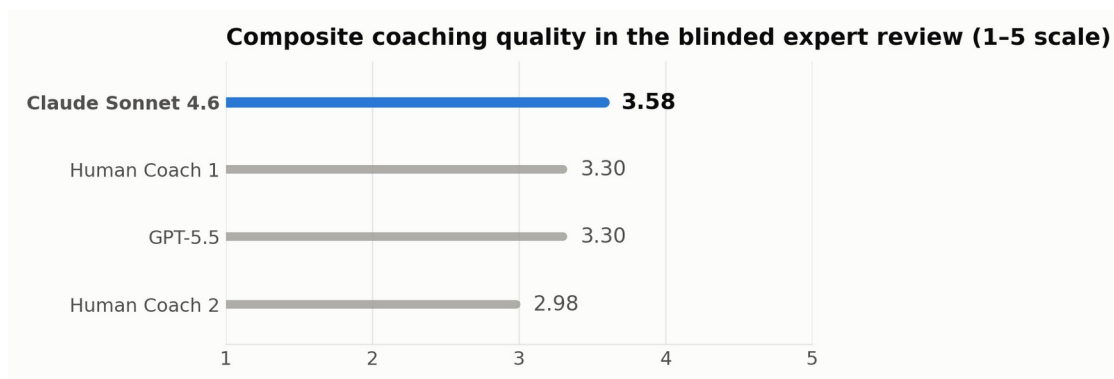


Figure 1. The result in one picture: mean composite quality (eight 1–5 dimensions) across 30 visits, scored by an independent expert blind to source identity and source type. Confidence intervals, ranks, and full detail in §5.

Why this study matters. The best evidence in community corrections says that what happens inside a supervision visit is the strongest lever an agency has on

participant outcomes. Labrecque, Viglione, and Caudy's (2023) meta-analysis of 25 officer-training studies found that officers who deliver core correctional practices with fidelity reduce violations and rearrests, while officers who drift produce outcomes no better than untrained peers. Chadwick et al. (2015) sizes the lever: untrained-officer caseloads recidivate at 49.9% versus 36.2% for trained ones — about 38% more often. The same literature is equally clear that the effect decays without reinforcement: fidelity after training, sustained by monitoring and coaching, is where the leverage lives.

Our companion paper (Green, Warren, & Meade, 2026) validated the missing measurement instrument — an AI grader that scores every recorded visit against a standards-based rubric about as reliably as expert human auditors, turning quality review from a staged annual sample into a census of real visits. But measurement alone does not change practice. A score without coaching is a dashboard; the literature's mechanism is feedback that helps the officer do something different in the next visit. The operational question this paper answers: **once RePath can measure visit quality at scale, can it responsibly act on that measurement — can an LLM generate officer-facing coaching that an expert would accept as equal to expert human coaching?**

The design in one paragraph. We took the same 30 supervision visits used in the grader validation and assembled four coaching outputs per visit — 120 total. Two sources were human: the authentic coaching comments that two of the three expert graders from the prior grading study wrote for these visits, de-identified throughout the results as **Human Coach 1 (HC1)** and **Human Coach 2 (HC2)**. Two sources were LLMs — Anthropic's Claude Sonnet 4.6 and OpenAI's GPT-5.5 — generating coaching from the transcript and score context only, with all human coaching text withheld from the generation inputs. A third expert, Valerie Meade, then reviewed all 120 outputs in a blinded workbook: outputs randomized within each visit, source identities *and* source types withheld until scoring was complete, and her own historical coaching for these visits excluded from the packet entirely, a fact stated to her explicitly. She scored each output on eight 1–5 dimensions (standards alignment, transcript grounding, prioritization, usefulness, tone, brevity, safety, and display fit), ranked the four outputs within each visit, and decided whether each could be shown next to a score in the product. Only after the completed workbook was returned did RePath Labs decode the internal source key.

The short answer. On this corpus, blinded expert review rated the best LLM coaching source as equal to or better than both directly compared expert human coaching sources. Claude Sonnet 4.6 earned the highest composite quality score (3.58 on the 1–5 index, vs. 3.30 and 2.98 for the two human sources and 3.30 for GPT-5.5), the best mean within-visit rank (2.07 of 4, with 10 first-place finishes — the most of any source), and the most display-ready outputs (13 of 30 rated suitable

to show next to a score as written). Against the stricter human comparator the difference is statistically significant after multiplicity correction (+0.60, Holm-adjusted $p = 0.006$); against the stronger human comparator the observed difference was not statistically significant (+0.28, n.s.), and a formal non-inferiority test makes the “equal to” half of the claim affirmative rather than an absence of evidence: at 95% confidence, the worst case against the study’s stronger human source is a 0.005-point deficit — a dead heat (§5.3). Pooled by type, the two LLM sources outscored the two human sources on the visit-level composite (3.44 vs. 3.14, paired difference +0.30, 95% CI +0.10 to +0.52, $p = 0.019$), with the LLM pool ahead on 20 of 30 visits. GPT-5.5 individually performed within the human range — not significantly different from the stronger human source and above the stricter one.

The numbers worth remembering.

- **120 coaching outputs, blind-scored.** 30 visits $\times$ 4 sources, randomized within visit; source identities and source types withheld until the completed workbook was returned; the reviewer’s own prior coaching excluded from the packet. The internal source key was stored separately and joined only after scoring was complete.
- **Composite quality (mean of eight 1–5 dimensions), by source:** Claude Sonnet 4.6 **3.58** (95% CI 3.41–3.74); Human Coach 1 3.30 (3.04–3.54); GPT-5.5 3.30 (3.20–3.41); Human Coach 2 2.98 (2.69–3.26). Confidence intervals are visit-level bootstrap intervals; the composite is a practical summary index, not a validated psychometric scale.
- **Within-visit rank (1 best, 4 lowest; lower is better), by source:** Claude 2.07 (best); HC1 2.37; GPT-5.5 2.73; HC2 2.83. Claude took rank 1 on 10 of 30 visits, the most of any single source. Comparing the two human outputs against the two LLM outputs within each visit, a human output was the single best in 16 of 30 visits — just over half — so expert human coaching at its best remains the thing to beat, visit by visit.
- **Pooled human vs. LLM:** composite 3.14 vs. 3.44, paired difference +0.30 (95% CI +0.10 to +0.52, Wilcoxon $p = 0.019$). The rank comparison points the same direction (2.60 vs. 2.40) but is not significant ($p = 0.20$). We report both so the pooled result cannot silently carry a claim the rank data does not.
- **Display readiness (the product bar):** rated as-written suitable to show next to a score: Claude 13/30, HC1 10/30, HC2 7/30, GPT-5.5 6/30. Adding “suitable after editing,” Claude reaches 24/30. The blinded reviewer applied the display bar without knowing which outputs were the product’s.
- **Failure modes are legible and different by type.** The dominant human-source failure was negative-only feedback without an anchoring example or a concrete alternative — 15 of the 16 lowest-scoring outputs were human-

source rows, with reviewer notes citing tone (“harsh,” “off-putting”) and actionability (“the officer would not know what to do next”). The dominant Claude failure was mis-prioritization: coaching that was accurate and well-toned but aimed at a compliance topic rather than the visit’s highest-value behavior-change opportunity — Claude scored 2 or lower on prioritization in 6 of 30 visits, though only one (visit 020) pulled its composite below 2.5. A secondary pattern in the reviewer’s notes is over-calibrated positivity: three Claude outputs were flagged as too positive relative to the concerns in the visit. GPT-5.5’s recurring note was brevity to the point of emptiness. These are exactly the failure modes a prompt revision and a delivery filter can target.

What this is for. Officer coaching: short, transcript-grounded, standards-linked guidance that helps an officer practice a specific behavior in the next visit, generated for every scored visit and delivered through a filtered cadence (immediate micro-coaching only for clear high-value opportunities, weekly recurring-theme digests, monthly trend review). The cadence is literature-informed design, not a validated finding of this study.

What this is not for. Hiring, firing, promotion, discipline, merit pay, or any other personnel decision. The validation does not authorize those uses.

How to read this paper. Sections 1–3, 6, and 9 are written for non-technical readers. Sections 4–5 and the appendices carry the methods, statistics, and full disclosures for technical reviewers. Each statistic gets a plain-language explanation the first time it appears.

2. Scope and intended use

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. It is not a supervisor memorandum, a performance review, or a discipline file.

Intended uses. Officer self-improvement (one transcript-grounded strength, one priority focus, one concrete next-visit behavior); supervisor coaching support (recurring themes, where human coaching is needed); agency quality improvement (aggregate themes, score movement, coaching volume at team or agency level); and product display (a short note beside a visit score or subsection, where expert review says it is display-ready).

Out-of-scope uses. Global judgments that an officer is good or bad; automated discipline or adverse employment action; using individual outputs as personnel metrics; and displaying every rubric item as a task after every visit.

The reliable unit of coaching is the priority, not the checklist. A single transcript may support many possible coaching points; the product should usually deliver one immediate priority. Broader patterns belong in a weekly digest or monthly trend review. This mirrors the companion paper’s finding that the reliable measurement unit is the visit score, not the individual item mark.

3. Background: from visit measurement to officer coaching

3.1 The evidence chain

This project’s claim structure is deliberately a chain, and each paper carries one link. The literature links officer fidelity to core correctional practices with better supervision outcomes and lower recidivism (Labrecque et al., 2023; Chadwick et al., 2015; Bonta et al., 2021). Agencies cannot act on fidelity at scale unless visit quality can be measured at scale; the companion grader paper validated that measurement link (Green, Warren, & Meade, 2026). This paper tests the next link: whether the measurements can be translated into expert-acceptable officer guidance. A third study — a consensus/adjudication study of the expert panel itself — will complete the reliability picture behind both papers.

The outcome literature matters because coaching is not an abstract feature. Chadwick et al.’s meta-analytic averages — roughly 36% recidivism for caseloads of trained officers versus roughly 50% for untrained — describe what is at stake when officer practice improves. Labrecque et al. locate the leverage precisely: fidelity *after* training is the moderator, the gap between high- and low-fidelity officers exceeding the gap between trained and untrained ones, and the authors call for “strategies to monitor and increase fidelity after training.” Measurement plus coaching is that strategy. None of this means the present study demonstrates a recidivism reduction; it means the study validates a mechanism the literature already treats as the one that matters.

3.2 Why rare human review is not enough

In most agency audit programs we have observed, quality review means one recorded visit per officer per quarter or year. The officer knows which visit counts, so the evaluated visit is closer to a performance than to normal practice; a single visit is also a sample of one. The same bottleneck constrains coaching: an officer whose work is reviewed once a year receives coaching once a year, attached to an unrepresentative visit.

The grader paper's census changes the unit of coaching. Instead of one observed visit as the coaching event, an officer's full body of recorded work becomes the coaching record. A note after one visit can say "try this next visit." A weekly or monthly trend can say "this appears to be a recurring focus" — and mean it, because the claim rests on all the visits rather than one. The product analogy we use internally: current systems are a coach seeing one random pitch a month; RePath is a coach seeing every live pitch thrown in actual games. The system must still not invent trend claims from a single visit, and the delivery design in §6.3 exists to enforce that.

3.3 What good coaching must do

The project's coaching-method research points to a narrow, practical standard drawn from two literatures. Audit-and-feedback and workplace-coaching research (Ivers et al., 2012, 2025; Kluger & DeNisi, 1996) governs form: compare performance to a clear standard, give task-specific feedback, preserve receptivity, avoid identity-level criticism, and identify one concrete next behavior — feedback that attacks the person rather than the task tends to backfire. Community-supervision practice research (Core Correctional Practices, EPICS/STICS, Motivational Interviewing; APPA, 2024; Bonta et al., 2021; Moyers et al., 2016) governs substance: what behaviors are worth coaching. The generation standard used in this study (§4.3) operationalizes both.

4. Methods

4.1 Foundation and corpus

The study uses the same 30-visit corpus as the grader validation: recorded community corrections supervision visits from multiple agencies, transcribed, independently scored by three expert graders during the prior study. The corpus is stratified across case types (10 Adult, 10 Juvenile, 10 Treatment Court) and supervision levels; full corpus characteristics are described in the companion paper. All 30 transcript IDs were mapped and included in the review package. The prior paper is methodological context, not a substitute: grading reliability and coaching usefulness are different questions, and the study design keeps the claims separate.

4.2 Coaching sources: two human, two LLM

Human sources. During the prior grading study, two of the three expert graders wrote coaching comments for each visit they scored. Those authentic comments — written by working experts in the ordinary course of review, complete for all 30 visits — are the human coaching sources, reported throughout under the de-

identified labels Human Coach 1 (HC1) and Human Coach 2 (HC2). Both graders are thanked by name in the Acknowledgements; individual results are reported only under the de-identified labels. They were **not** collected under a prospective RePath coaching protocol, and we do not present them as standardized product copy; the comparison is against real expert coaching as it actually occurs, with the format consequences of that choice disclosed in §5.5 and §7.

LLM sources. Two models generated coaching for all 30 visits: Anthropic’s Claude Sonnet 4.6 (model identifier claude-sonnet-4-6) and OpenAI’s GPT-5.5. Generation inputs were the visit transcript and available score/rubric context only. **Human coaching comments were withheld from the generation manifest**, so the models could not copy or paraphrase the human text. Thirteen visits carried full item-level score context; seventeen carried overall/tier scores only. Two initially truncated Claude responses (visits 011, 012) were regenerated with a larger output budget; all 60 LLM outputs were valid before packet assembly. A sensitivity analysis excluding these two visits leaves every §5 conclusion unchanged (pooled LLM-human +0.27, $p = 0.042$; Claude remains the highest-rated source and non-inferior to Human Coach 1 at the 0.4 margin, $p = 0.0006$).

Excluded source. Valerie Meade also wrote coaching comments during the prior study (29 of 30 visits). Because she is this study’s expert reviewer, her coaching was excluded from the packet entirely. The consequence for the claim boundary is stated in §7: this study directly compares two human coaching sources; it does not rank AI coaching against the evaluator’s own historical coaching.

4.3 Coaching design standard

The LLM generation prompt implements a pre-specified design standard: identify a transcript-grounded strength; select one priority improvement focus; name the supervision-standard domain; explain why the behavior matters; give one concrete next-visit behavior; avoid unsupported trend claims, disciplinary language, diagnosis, and exaggerated conclusions. The standard exists so that “good coaching” is defined before, not after, the review.

4.4 The blinded expert review

Reviewer. Valerie Meade, Deputy Director at the Crime and Justice Institute, with nearly three decades in adult and youth justice, including leading the fidelity team at the Johnson County, Kansas Department of Corrections — direct professional experience in transcript-based scoring, one-on-one officer coaching, and rubric design. She also trains trainers in evidence-based practices and provides fidelity coaching to agencies nationwide.

Packet. One workbook, one instruction document, and the 30 transcripts. The workbook presented 120 rows — four coaching outputs per visit, labeled only Coach A–D, **randomized within each visit** so that no label consistently identified a source.

Blinding. Source identities and source types were withheld from all reviewer-facing materials until scoring was complete. The instructions told the reviewer only that outputs came from “the coaching comparison set”; they did not disclose that models were involved at all, which models, or which humans. The instructions also stated explicitly that none of the outputs were written by her or attributed to her, so the review remained independent without her needing to guess. The internal source key (visit × label × true source) was stored separately in a restricted folder and was not joined to the workbook until the scored workbook was returned in August 2026. One honest qualifier: because outputs were reviewed as originally written or generated, some source-type inference from style, length, or formatting may have been possible; §7 treats this as a limitation, and §5.5 explains why it is also partly the product question itself.

Instrument. For each output: eight 1–5 dimensions (standards alignment; transcript grounding; prioritization; usefulness/actionability; tone/receptivity; brevity/cognitive load; safety/overclaiming; subsection/display fit), each with anchored descriptions at every scale point (1–5); a within-visit rank (1 best to 4 lowest, no ties); and a product-display decision (“Would show next to score?": yes / edit first / digest only / do not show). The display lens instructed the reviewer not to re-grade the visit but to judge whether the coaching translates a score gap into one accurate, brief, non-disciplinary, officer-facing next action.

4.5 Statistics

Primary outcome. The within-visit composite: the mean of the eight dimension scores for each output, compared across sources and pooled by type (the two humans versus the two LLMs, averaged within visit). We treat the composite as a practical summary index, not a validated psychometric scale, and report every dimension separately so the composite cannot hide a weak component.

Secondary outcomes. Within-visit rank; each dimension; the display decision; and qualitative failure modes from the reviewer’s free-text notes.

Inference. The visit is the independent unit. Confidence intervals are nonparametric bootstrap intervals resampling visits (10,000 draws). Paired comparisons use the Wilcoxon signed-rank test on visit-level values. The six pairwise source comparisons are Holm-adjusted for multiplicity; the pooled human-versus-LLM composite comparison is the single pre-specified primary test. Analyses were scripted (`analyze.py`) and run only after the completed workbook

was returned and joined to the source key; the preliminary decoded summary was reproduced exactly by the final script before any drafting of §5.

Supplementary non-inferiority analysis. Because the study’s central claim is “equal to or better,” difference tests alone are not enough: a non-significant difference is an absence of evidence, not evidence of equality. We therefore supplement the Wilcoxon tests with one-sided non-inferiority tests on the visit-level paired composite differences. The null hypothesis is that the LLM source is worse than the comparator by at least 0.4 points on the 1–5 composite — a margin well under one rubric anchor step and smaller than the smallest gap the reviewer’s notes treated as meaningful — at one-sided $\alpha = 0.05$. In honest disclosure: the margin was set before these tests were computed but after the primary decoding; the follow-up study will pre-register it.

Data completeness. The returned workbook is complete: all 120 rows carry a within-visit rank, all eight criterion scores, and a display decision. The review was completed in full before any results or source information were shared with the reviewer, and all analyses use the complete 120-row dataset.

4.6 Internal QA dry run (readiness only)

Before the packet was sent, an internal GPT-5.5 extra-high review scored all 120 rows as a workability check on the workbook, rubric anchors, rank fields, and display fields, without access to the source key during scoring. That dry run confirmed the packet was scorable. It is readiness evidence only; no dry-run number is validation evidence, and none appears in §5.

5. Results

5.1 Headline: blinded expert review by source

Valerie Meade returned the completed workbook in August 2026. Decoded against the internal source key, the four sources performed as follows.

Source	Type	Composite (95% CI)	Median	Mean rank	Rank 1/2/3/4	Display Y/E/D/N
Claude Sonnet 4.6	LLM	3.58 (3.41–3.74)	3.62	2.07	10/12/4/4	13/11/2/4
Human Coach 1	Human	3.30 (3.04–3.54)	3.25	2.37	9/6/10/5	10/16/0/4
GPT-5.5	LLM	3.30 (3.20–3.41)	3.25	2.73	4/9/8/9	6/15/4/5
Human Coach 2	Human	2.98 (2.69–3.26)	3.00	2.83	7/3/8/12	7/13/0/10

Table 1. Blinded expert review results by decoded source. Composite = mean of the eight 1–5 dimensions ($n = 30$ visits per source, complete data). CIs are visit-level bootstrap intervals. Median is the median visit-level composite. Rank: how often the source’s output placed 1st/2nd/3rd/4th among the four outputs for a visit (1 best, no ties, all 30 visits). Display Y/E/D/N is the reviewer’s product decision for each output – “Would show next to score?": **Yes**, show as written; **Edit first**, useful but needs editing before display; **Digest only**, hold for a weekly summary rather than immediate display; **No**, do not show next to a score.

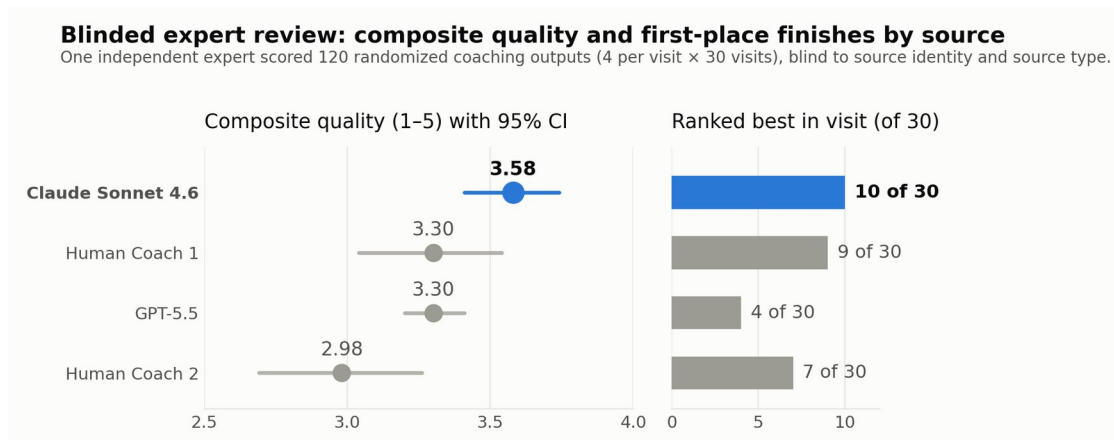


Figure 2. The headline result at a glance: highest composite quality (left, with 95% bootstrap confidence intervals) and the most first-place finishes within visits (right). Sources are ordered by composite score; the two human sources are de-identified as in Table 1; display decisions are reported in Table 1 and §5.5.

Claude Sonnet 4.6 leads on every summary measure: highest composite, best mean rank, most first-place finishes, most display-ready outputs. Human Coach 1 is the stronger human source and sits statistically level with Claude. GPT-5.5 lands inside the human range – not significantly different from HC1, above HC2. Human Coach 2’s comments, scored blind, trail on the product-oriented dimensions; §5.4 shows why, and §5.5 explains why that says less about her expert judgment than it does about raw grader notes as product copy.

5.2 The pre-specified primary comparison: pooled human vs. LLM

Pooling the two sources of each type within visit, LLM coaching scored 3.44 and human coaching 3.14 on the composite – a paired difference of **+0.30** (95% CI +0.10 to +0.52; Wilcoxon signed-rank $p = 0.019$), with the LLM pool higher on 20 of 30 visits. In plain terms: on two of every three visits, the blinded expert scored the model coaching higher than the expert human coaching for that same visit.

The rank view is directionally consistent but weaker: pooled mean rank 2.40 (LLM) versus 2.60 (human), not statistically significant ($p = 0.20$), and a human output was

still the single best output on 16 of 30 visits — just over half. Both views are reported deliberately. The composite says the LLM pool is scored higher on average; the rank data says expert human coaching at its best still wins the visit about half the time. A trustworthy summary is “at least as good, often better” — not “humans obsolete.”

5.3 Pairwise source comparisons

Six paired Wilcoxon tests on visit-level composites, Holm-adjusted:

Comparison	n	Means	Difference	Holm-adj. p
Claude vs. HC2	30	3.58 vs. 2.98	+0.60	0.006
Claude vs. GPT-5.5	30	3.58 vs. 3.30	+0.27	0.065
GPT-5.5 vs. HC2	30	3.30 vs. 2.98	+0.33	0.23
Claude vs. HC1	30	3.58 vs. 3.30	+0.28	0.32
HC1 vs. HC2	30	3.30 vs. 2.98	+0.32	0.32
GPT-5.5 vs. HC1	30	3.30 vs. 3.30	+0.01	0.89

Table 2. Pairwise composites (paired by visit, all 30 visits). Only Claude vs. HC2 survives multiplicity correction.

Head-to-head by visit, Claude’s composite was at least as high as HC1’s on 20 of 30 visits and at least as high as HC2’s on 23 of 30; Claude out-ranked HC1 on 18 of 30 and HC2 on 21 of 30. This is the evidentiary basis for the paper’s central claim, stated in the form the analysis supports: **on this corpus, an independent expert reviewing randomized, source-blinded coaching rated at least one RePath LLM source as equal to (vs. HC1: +0.28, n.s.) or better than (vs. HC2: +0.60, Holm-adjusted $p = 0.006$) the two directly compared expert human coaching sources on the pre-specified quality and product-usability measures.** The two human sources did not differ significantly from each other (+0.32, n.s.), which is itself consistent with the companion paper’s finding that expert judgments on this material vary.

The “equal to” half of the claim is affirmative, not just an absence of difference. Supplementary non-inferiority tests (§4.5; margin 0.4 points) reject the hypothesis that Claude is meaningfully worse than either human source: against Human Coach 1, the one-sided 95% lower confidence bound on the difference is -0.005 points ($p = 0.0002$; non-inferior even at a stricter 0.25 margin, $p = 0.002$) — in plain terms, at 95% confidence the worst case against the study’s stronger human source is a dead heat. The conclusion is insensitive to the margin choice: with a lower bound of -0.005 , non-inferiority holds at any margin larger than 0.005 points — the data rule out anything worse than a near-exact tie. Against Human Coach 2 the lower bound is $+0.35$, and for the pooled LLM-versus-human comparison it is $+0.12$ — both bounds above zero, meaning those comparisons establish superiority, not merely

non-inferiority. Together with the difference tests, this is the statistical basis for stating that the strongest LLM source performed at or above the level of the directly compared expert human coaching sources on this corpus.

5.4 Dimension profile: where each source wins

Dimension	Claude	GPT-5.5	HC1	HC2
Standards alignment	3.13	2.63	3.07	2.97
Transcript grounding	3.40	2.90	3.07	2.67
Prioritization	3.17	2.73	3.40	3.23
Usefulness/actionability	3.60	3.30	2.73	2.47
Tone/receptivity	4.13	3.97	3.23	2.47
Brevity/cognitive load	3.87	4.03	4.10	4.07
Safety/overclaiming	3.77	3.67	3.50	3.07
Subsection/display fit	3.53	3.20	3.27	2.87

Table 3. Mean dimension scores (1–5), all 30 visits per source. Bold = best source per dimension.

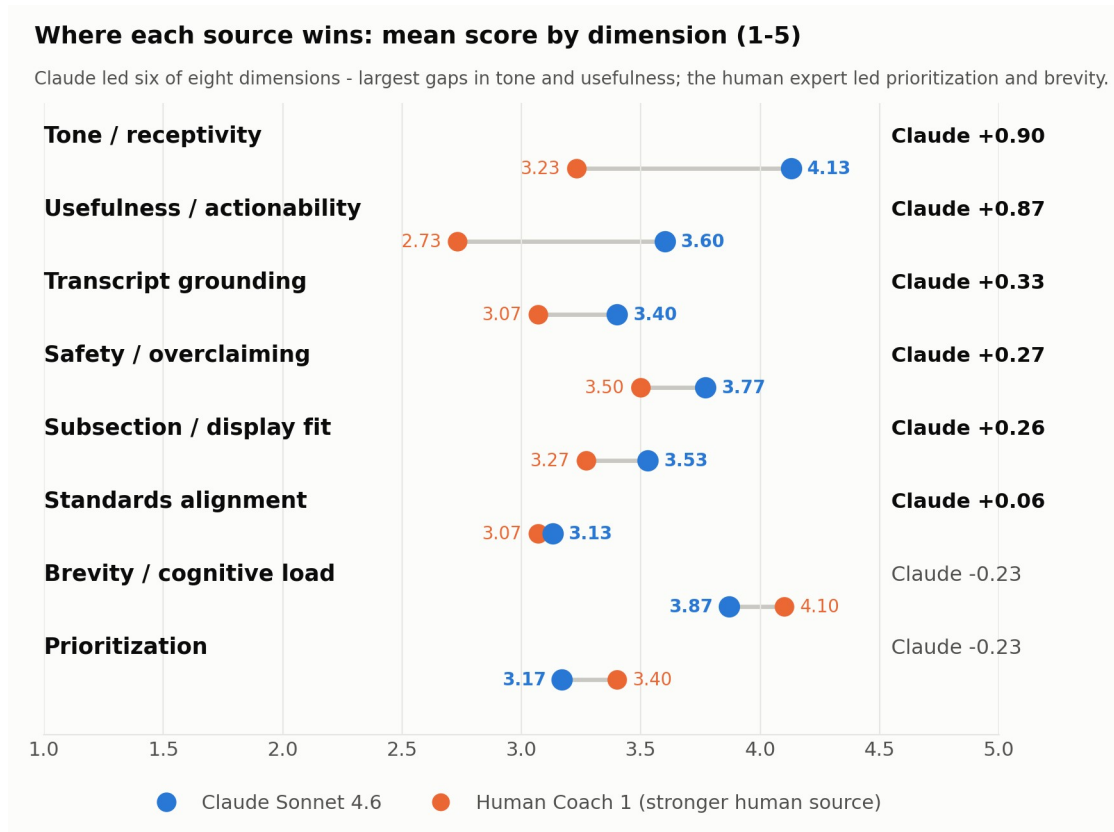


Figure 3. The dimension profile behind Table 3, comparing the strongest LLM source against the stronger human source. Claude led six of eight dimensions, with the largest gaps on tone/receptivity (+0.90) and usefulness/actionability (+0.87); Human

Coach 1 led prioritization and brevity – the pattern discussed below and the basis for the round-2 improvement target. Full four-source detail in Table 3.

The pattern is coherent and, we think, the study's most useful product finding. The LLM advantage concentrates where delivery quality lives: usefulness (+0.87 over the better human), tone (+0.90), grounding, safety, and display fit. The human advantage concentrates where expert judgment lives: **prioritization** – Human Coach 1 was the best in the study at picking the highest-value coaching opportunity (3.40), and both humans beat GPT-5.5 there. Brevity is a wash. The blinded expert, in other words, found that the experts know best *what* to coach, and the strongest model was better at *how* to say it – while Claude's prioritization (3.17) held close enough to the human standard for the overall package to win.

5.5 Display readiness and failure modes

Display readiness. Rated suitable to show next to a score as written: Claude 13/30, HC1 10/30, HC2 7/30, GPT-5.5 6/30; adding “suitable after editing,” Claude reaches 24/30. The highest “do not show” rate belonged to a human source (HC2, 10 of 30). The blinded reviewer was applying RePath's own product bar without knowing which outputs were the product's – a design feature worth keeping in future rounds.

Human-source failure mode: unanchored negativity. Fifteen of the sixteen lowest-scoring outputs (composite ≤ 2.5) were human rows, and the reviewer's notes are consistent: feedback that lists what the officer did wrong without an anchoring example or a concrete alternative (“the officer would have a very difficult time knowing what to do next”; “negative, doesn't provide examples to anchor”). This should be read fairly: these were working grader notes never intended as officer-facing product copy, and the graders were never asked to write to the display standard. The finding is not that the experts coach badly; it is that authentic expert notes, as they occur in real QA workflows, are usually not deliverable to officers without editing – which is precisely the gap the product proposes to fill.

Why the non-standardized comparison is also the right comparison. A reviewer might object that the human sources were “not trying” to write product coaching, so the comparison flatters the LLM. The objection assumes a standardized human coaching practice exists to compare against. It does not. In the field, officer coaching – where it exists at all – has no common format, cadence, or content standard: it arrives as sparse, unstructured supervisor or QA notes attached to the occasional reviewed visit, written exactly the way the comments in this study were written. Where richer human coaching does occur, it is often delivered verbally in supervision meetings – feedback that is rarely written down, remembered consistently, or tracked over time, and that reaches only the officers a supervisor has time to sit with. The authentic grader notes are therefore not a weak stand-in

for real coaching; they are representative of the coaching officers actually receive when they receive any, and they were produced by experts engaged specifically to review these visits — a level of expert attention most officers never get. Two readings of the result should be kept separate. As a statement about expert *capability*, the comparison is format-confounded and §7 says so. As a statement about what an officer would actually receive — the operational question the product answers — the comparison is ecologically valid — and, since most officers receive no per-visit coaching at all, arguably a harder test than the field baseline. A prospectively standardized human arm remains worth running (§8), but it would test a comparator that does not exist in practice today.

Claude failure mode: off-priority focus. Claude’s only low-composite output (visit 020; it drew “do not show” on four visits in total) was coaching that was positive, grounded, and well-written — about the wrong thing, reinforcing compliance-related goals rather than the visit’s behavior-change opportunity. Visit 020 was the extreme case, not the only one: Claude scored 2 or lower on the prioritization dimension in 6 of 30 visits. The reviewer’s notes also flag a secondary pattern of over-calibrated positivity — three Claude outputs (visits 089, 113, 114) were rated too positive relative to the concerns caught in the same visits. The same note appears in milder form on visit 065 (“accurate, but ignores the most important missed opportunity”). Prioritization is the dimension to target in the next prompt revision, and the human graders’ strength there suggests the training signal.

GPT-5.5 failure mode: brevity past the point of usefulness. Recurring notes praise the organization and labeling but find the content “too brief to give meaningful feedback” and occasionally aimed at a secondary issue. Its composite is held up by tone and brevity scores; its standards alignment (2.63) and prioritization (2.73) are the weakest in the study. A favorable pooled LLM result must not hide this: one model is doing most of the work, and model choice matters as much here as it did in the grader paper’s prompt-model pairing finding.

What “blind” bought us. Because sources and source types were withheld and randomized, these patterns cannot be an artifact of the reviewer knowing which outputs to favor. The caveat runs the other way: outputs were reviewed as originally written, so style may have signaled source type even without labels. The defense against over-reading is in the design — the dimensions specifically separate content judgments (standards, grounding, prioritization) from format judgments (brevity, display fit), and the LLM advantage on content dimensions like grounding survives that separation — but format-driven inference cannot be fully excluded, and we say so plainly in §7.

6. Discussion

6.1 What this study establishes

Combined with the companion grader paper, the evidence chain now reads: (1) the literature ties officer fidelity to core correctional practices to materially lower recidivism; (2) an LLM can measure visit-level fidelity about as reliably as expert human auditors; and now (3) an LLM can convert those measurements into officer-facing coaching that a blinded independent expert scores as equal to or better than the authentic expert human coaching it was compared against, and more often display-ready. The middle of the product thesis — *agencies can responsibly act on validated visit scores* — has moved from plausible to evidenced, within the boundaries stated below.

The operational framing matters as much as the statistics. Human coaching of this quality currently reaches an officer only when an expert reviews a visit — for most officers, a handful of times a year at best. The study’s comparison sources are what expert review actually produces. A system that generates coaching of blinded-expert-comparable quality for every scored visit — delivered under the expert-calibrated display filter §9 prescribes — does not need to beat human experts to change practice; it needs to match them and be present, at a coverage level no human program can staff.

6.2 What this study does not establish

It does not show that officers read, accept, or act on the coaching; that practice fidelity improves; that participant outcomes improve; or that recidivism falls. Those are downstream questions requiring field studies, and the recidivism numbers cited in this paper are the literature’s, not ours. It does not validate any specific delivery cadence. It does not compare the LLMs against the evaluator’s own coaching (see §7). And it does not establish that *any* LLM coaches well: one of the two models tested merely reached the human range, and the grader paper’s lesson — performance is a property of the prompt-model pairing, revalidated per version — applies with full force to coaching.

6.3 How coaching should be delivered

The delivery design remains the literature-informed standard described in §2 and §3 — immediate micro-coaching for clear high-value opportunities, a weekly digest for recurring themes, and monthly trend review — now with one empirical anchor: the blinded display decisions. Generate coaching internally for every scored visit; show immediate micro-coaching only when the transcript supports a clear, high-value opportunity (the reviewer’s “yes” bar — met by 13 of 30 Claude outputs as written); route recurring themes to a weekly digest (her “digest only” lane); require editing where she would edit (11 of 30); and suppress the rest. Monthly trend review

carries score movement and repeated domains. Supervisor and agency views stay aggregate and developmental; escalation stays reserved for safety and legal-compliance issues. None of this cadence is proven by this study; what the study adds is an expert-calibrated filter for the first delivery tier.

6.4 Product and engineering implications

Three concrete directions fall out of §5. First, target prioritization: the next prompt revision should train on the reviewer's rank-1 examples and Human Coach 1's prioritization pattern, since picking the right opportunity is now the measured gap. Second, keep the display filter expert-calibrated: the 120 display decisions are a labeled dataset for the delivery classifier. Third, preserve generated-versus-delivered separation in the data model (visit ID, rubric and prompt versions, model ID, evidence excerpt, delivery level, validation status, not-for-discipline flag), so the agency can always audit what was generated, what was shown, and why.

7. Limitations

- **One expert reviewer.** All 120 judgments come from a single independent expert. She is well qualified and was blinded to source identity and source type, but single-reviewer designs cannot separate her calibration from the construct.
- **Two directly compared human sources.** The study directly compares two human coaching sources; a third human expert served as evaluator. It does not directly rank AI coaching against that evaluator's own historical coaching. A claim that the LLM matched or beat *all three* human coaching sources would require a second reviewer scoring her coaching, and is not made here.
- **One generation per model per visit.** Each LLM contributed one generation per visit, matching production behavior; the study evaluates the delivered outputs, not run-to-run stability, which the follow-up study can sample directly.
- **Human comparison text was not prospectively standardized.** The human coaches' comments are authentic prior grader notes, not product-formatted coaching written to the display standard. Part of the LLM advantage on format-adjacent dimensions (brevity, display fit, tone) plausibly reflects that difference. The content dimensions partially control for it; they do not eliminate it. §5.5 explains why the comparison is nonetheless the ecologically valid one: field coaching has no common, consistent standard, and unstandardized expert notes are what real coaching looks like where it exists at all.

- **Source-type inference was possible.** Identities and types were withheld and order randomized, but outputs were reviewed as written; style, structure, or polish may have signaled type. Perfect blinding of type is claimed nowhere in this paper.
 - **30 visits, one corpus.** Sufficient for a first blinded expert review; not a multi-site field validation. The corpus also carries the prior study's mix of agencies and visit types.
 - **Mixed score-context depth.** Thirteen visits carried item-level score context into generation; seventeen carried tier/overall context only. We did not power a subgroup analysis on this split.
 - **Composite is a practical index.** Equal-weighted mean of eight dimensions, chosen before decoding but not psychometrically validated; every dimension is therefore reported separately.
 - **Post hoc choices and conflicts.** RePath Labs designed, funded, and analyzed the study of its own product. The primary analysis, the claim boundary, and the required limitation language were specified in writing before the final analysis was run; in full disclosure, that memo already contained a preliminary source-level decode (flagged as a verification target only), so the analysis plan cannot be called fully pre-registered. The final script reproduced the preliminary verification targets exactly, the rubric and blinding design predate any decoding, and the follow-up study will be pre-registered — but independent replication is the real control.
 - **Product formatting may favor LLM outputs.** The display-fit criterion asks whether an output can sit next to a score in RePath's product; outputs generated to RePath's design standard have a structural head start on that question. The rubric's content dimensions and the rank measure temper, but do not remove, this.
-

8. Future work

1. **The agreement/adjudication study** (in preparation): consensus references and inter-rater structure for the expert panel, completing the reliability foundation under both papers.
2. **Coaching validation round 2 (in preparation).** A pre-registered follow-up study designed to improve the LLM coaching and test it under stricter conditions. First, the coaching prompt is revised using expert human input from this study — the blinded reviewer's top-ranked exemplars, her display decisions, and her prioritization notes, since prioritization is the measured gap in §5.4. The improved coaching is then generated for a held-out set of visits that no prompt in this study has seen, alongside the unrevised round-1 prompt and expert human coaching written for those same visits, and all

outputs are blind-reviewed under the same rubric — with the analysis, non-inferiority and superiority margins, and a blinding check pre-registered before any review begins.

3. **Longitudinal product measurement:** acknowledgment rates, repeated-focus improvement, score movement.
 4. **Field outcome studies** linking coached officers to fidelity and, eventually, participant outcomes — the final link in the chain.
-

9. Responsible use and recommendations for agencies

Treat coaching as quality improvement, never discipline; the tool's data should not feed personnel decisions, and product copy, contracts, and policy should say so. Require transcript evidence for every note an officer sees. Deliver one priority behavior per note with a concrete next-visit action. Filter delivery (immediate / digest / trend) rather than flooding officers; keep supervisors in the loop for trends, sensitive visits, and training plans. Track acknowledgment, helpfulness, edits, suppression, and escalation. Revalidate prompts whenever models, rubrics, or display formats change. And preserve the boundary this paper preserves: coaching quality at the level of the directly compared expert sources is supported here; behavior change and outcome impact are not yet.

Data and ethics

The visit transcripts are the same 30 de-identified transcripts used in the companion grader study: 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 human coaching comments evaluated here were written by the companion study's expert graders on that de-identified corpus; both consented to the use of their comments and to being thanked by name. The external expert reviewer was compensated for her time, reviewed only de-identified text, was blinded to source identity and source type, and consented to be named.

Competing interests and funding

RePath Labs designed, funded, and conducted this study and has a commercial interest in officer-visit grading and coaching tools. Valerie Meade served as the independent expert reviewer in a compensated consulting role; her review was conducted blind to source identity and source type, and her own prior coaching was excluded from the reviewed packet.

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 blinded expert review of all 120 coaching outputs. She is credited as a contributor rather than an author deliberately: because her blinded 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 graders **Maureen Flanagan** and **Katie Myers** are acknowledged as the human coaching sources; their comments were written during the prior grading study and used here with the study's disclosures.

Acknowledgements

The authors thank Valerie Meade for performing the blinded expert review at the center of this study, and expert graders Maureen Flanagan and Katie Myers, whose coaching comments from the prior grading study served as the human comparison sources. All three experts also served as graders in the companion grader-validation study; their sustained, careful work underlies both papers.

References

- American Probation and Parole Association [APPA] (2024). National Standards for Community Supervision. https://www.appa-net.org/eweb/docs/APPA/National_Standards_Community_Supervision_FINAL.pdf
- 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. DOI: 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. DOI: 10.1177/0093854815595661.
- Green, T., Warren, M., & Meade, V. (2026). Validating an AI Grader for Community Corrections Supervision Visits: A reliability and agreement study against expert human auditors. RePath Labs working paper. DOI: 10.31219/osf.io/9wq32_v1.

- Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., O'Brien, M. A., Johansen, M., Grimshaw, J., & Oxman, A. D. (2012). Audit and feedback: Effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews*, (6), CD000259. DOI: 10.1002/14651858.CD000259.pub3.
- Ivers, N., Yogasingam, S., Lacroix, M., Brown, K. A., Antony, J., ... Grimshaw, J. M. (2025). Audit and feedback: Effects on professional practice. *Cochrane Database of Systematic Reviews*, 3, CD000259. DOI: 10.1002/14651858.CD000259.pub4.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. DOI: 10.1037/0033-2909.119.2.254.
- 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. DOI: 10.1080/07418825.2022.2120062.
- Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The Motivational Interviewing Treatment Integrity code (MITI 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65, 36–42.

Appendices

Appendix A: Review packet inventory

Item	Description
Reviewer packet	One review workbook, one instruction document, and 30 visit transcripts, sent to the reviewer as a single package (July 2026)
Reviewer instructions	Blinding statement, anchored scoring rubric, display lens, and return instructions
Review workbook	120 rows — four blinded coach outputs per visit, labeled Coach A–D and randomized within visit; returned complete, August 2026
Transcripts	30 de-identified visit transcript files
Internal source key	Visit × label × true source mapping; stored separately from all reviewer-facing materials and joined only after the scored workbook was returned
Analysis package	Reproducible analysis script (<code>analyze.py</code>), decoded row-level table, analysis report, and claims ledger (see Appendix D and Materials availability)

Appendix B: Blinding statement (as given to the reviewer)

“Each visit has four blinded coach outputs from the coaching comparison set. Source identities and source types are intentionally withheld until scoring is complete. None of the coaching outputs you will review were written by you or attributed to you. Any prior comments from you were excluded so the review remains independent. ... Return the completed workbook only. Do not ask for or open the source key until after scoring is complete.”

Appendix C: Source identities (decoded post-return)

Human Coach 1 (human expert grader), Human Coach 2 (human expert grader), RePath AI – GPT-5.5 (LLM), and RePath AI – Claude Sonnet 4.6 via Anthropic (LLM). Labels A–D were randomized within each visit; no label mapped consistently to a source. The two human sources are de-identified in this paper; both graders are named in the Acknowledgements.

Appendix D: Reproducing the numbers

Every statistic in §5 is computed by `analyze.py` from the immutable copy of the returned workbook joined to the internal source key: composites from the eight dimension columns (complete rows only), bootstrap CIs resampling visits (10,000 draws, seed 42), Wilcoxon signed-rank tests on visit-level paired values, Holm adjustment across the six pairwise tests. The preliminary decoded table was reproduced exactly before Results were written. The decoded row-level table (`decoded-source-table.csv`), analysis report, and claims ledger are archived with the frozen study record and may be made available to interested parties by agreement (see Materials availability). Model versions: `claude-sonnet-4-6` (Anthropic API) and `GPT-5.5` (OpenAI API); one generation per model per visit.