

Establishing and Meeting a Reference Standard for AI Grading of Community Corrections Supervision Visits

An expert adjudication and multi-model agreement study

Ted Green and Madeline Warren (RePath Labs); Valerie Meade (consultant)

Expert graders and adjudication panel: Valerie Meade, Maureen Flanagan, and Katie Myers

Version 1.0 · September 30, 2026

OSF version 1 DOI: https://doi.org/10.31219/osf.io/xk5u9_v1

Contents

Abstract.....	2
1 Executive summary.....	3
2 Scope and intended use.....	4
3 Background.....	5
4 Methods.....	7
5 Results.....	14
6 Discussion.....	20
7 Limitations.....	24
8 Future work.....	25
9 Responsible use and recommendations for agencies.....	25
10 Conclusion.....	26
Data and ethics.....	26
Competing interests and funding.....	26
Author contributions.....	27
Acknowledgements.....	27
References.....	27
Appendix A Measurement definitions.....	29
Appendix B Detailed Astra results.....	31
Appendix C Model workflows and execution.....	32
Appendix D Operational policy boundaries.....	34

Working paper (RePath Labs). This preprint has not undergone journal peer review and is available for comment.

Suggested citation: Green, T., Warren, M., & Meade, V. (2026). 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.

Companion papers: Paper 1, the earlier AI-grading study (Green, Warren, & Meade, 2026), https://doi.org/10.31219/osf.io/9wq32_v1; Paper 2, the coaching-validation study (Green & Warren, 2026), https://doi.org/10.31219/osf.io/m7d5g_v1. Full citations in References.

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

ORCID iDs: Ted Green 0009-0008-9264-6568; Madeline Warren 0009-0000-0141-8810; Valerie Meade 0009-0002-7261-0601.

License: © 2026 RePath Labs. All rights reserved.

Materials availability. The rubric, adjudicated reference labels, human and model marks, scoring code, versioned grading policies, and model-version details are described in Appendices A–D; the full verification package may be made available to interested parties by agreement. The proprietary grader instructions and the PII-bearing transcripts are not distributed.

Scope. The grading workflow evaluated here is intended to support review and coaching of officer practice. It is not intended for hiring, firing, promotion, discipline, or any other personnel decision.

Abstract

Limited review capacity and differences in expert judgment make consistent assessment of supervision practice in community corrections difficult. In our earlier grading study, three experienced expert graders assigned substantially different scores to the same visits. Here, we establish a shared expert reference and test whether AI grading can reproduce it, providing a means to extend reference-grounded review beyond manual audits. The three experts jointly adjudicated 43 rubric items across ten visits, five adult and five juvenile, reaching unanimous decisions on 429 of 430 items and documenting their reasoning. We used this reference to develop general grading instructions. The selected GPT-6 Astra workflow, combining three runs by item-level majority vote, matched 387 of 429 decisions (90.21%; Cohen's kappa 0.832), with mean absolute visit-score error of 4.57 points under the rubric's original tier weights. It exceeded all three experts' initial independent agreement rates (49.65%–86.01%; mean 67.06%) and was 23.15 percentage points above their mean. It met all four preset development targets; four other tested workflows reached 80.89%–85.55% agreement. All ten visits were used in development, so performance on new visits remains to be measured. These results provide a foundation for reviewing more visits against a shared standard and using those grades to guide officer coaching.

1 Executive summary

Why this study matters. Spot audits by different reviewers do not, by themselves, establish a shared scoring interpretation or provide review of every visit. In Paper 1, three experienced expert graders' weighted scores for the same visits differed by 19.0 to 33.3 points on average (Green, Warren, & Meade, 2026). This study addresses that consistency problem by establishing a shared expert reference and testing an automated workflow that can apply it more broadly. Paper 2's evaluation of AI-generated coaching supplies the companion evidence for turning visit review into useful feedback.

The short answer. Experts established a shared reference, and AI reproduced it with strong agreement on the ten visits studied. The panel reached unanimous decisions on 429 of 430 rubric items. Against that reference, the experts' initial independent ratings agreed on 49.65%–86.01% of decisions (mean 67.06%). The selected GPT-6 Astra workflow agreed on 90.21% (387/429), exceeding all three initial ratings and standing 23.15 percentage points above their mean. Cohen's kappa was 0.832, a measure of agreement adjusted for chance, and average absolute visit-score error was 4.57 points. The workflow met all four preset development targets.

The design. Ten visits, five adult and five juvenile, were drawn from Paper 1's 30-visit corpus, selected to include human scoring disagreements, inconsistent AI grading, and comparison cases with greater agreement or higher-quality supervision. We used the adjudicated reference to develop general grading instructions; grading calls never received the reference labels. All ten visits were used for development, so accuracy on new visits remains to be measured.

Three findings. Tier agreement was 100.00% for Foundations, 94.29% for Structure, and 87.70% for Core Correctional Practice Execution, achieved without changing the rubric's original score weights. The four other tested workflows reached 80.89%–85.55%, showing why model, policy, and workflow must be tested together. Most remaining errors were shared by all three Astra runs, so review should also sample items the runs agree on, especially in modeling, problem solving, and role clarification.

What this is for. Developmental review of officer practice: identifying evidence, pointing to an omitted step, and supplying structured context for coaching. A

reliable grading layer could extend review well beyond the small share of visits an agency can audit by hand.

What this is not for. Hiring, firing, promotion, discipline, merit pay, or any other personnel decision. Nor does this study measure coaching quality, changes in officer practice, or recidivism.

2 Scope and intended use

We evaluated transcript-based grading of community corrections supervision visits. The unit of grading was a rubric item within a visit. Each item received YES, NO, or NA. YES means the specified behavior or step is evidenced under the operational scoring policy; NO means an applicable criterion is not evidenced; NA means the criterion does not apply under the opportunity rules. NA is a substantive label, not missing data.

We applied Paper 1's 43-item rubric (instrument edition 2.0, archived for that study as v1.0). It has five items in Tier 1 (Foundations), seven in Tier 2 (Structure), and 31 in Tier 3 (Core Correctional Practice Execution). Tier 3 covers authority, modeling, reinforcement, disapproval, problem solving, role clarification, cognitive restructuring, active listening, and feedback. We retained the original tier weights of 20%, 30%, and 50%; practice execution therefore contributes half of the weighted visit score.

The intended use is to guide review and coaching of officer practice. This study does not validate use for participant risk prediction, decisions about liberty, automatic sanctions, employment discipline, or a general determination of officer competence. A transcript also does not capture everything: tone, nonverbal conduct, recording failures, and work outside the recorded encounter may matter.

In this paper, *reference standard* means the documented decisions the expert panel reached through adjudication for this corpus and rubric version. It does not mean every reasonable expert must make the same judgment, that the reference is error-free, or that an item label proves the underlying intervention worked. Agreement with the standard is the measured outcome. Better supervision and better participant outcomes remain the purpose of the broader program.

3 Background

3.1 Why grading fidelity matters

Officer training and implementation fidelity are tied to the outcomes the field cares about. Chadwick, DeWolf, and Serin's (2015) meta-analysis of ten studies reported weighted recidivism of 36.2% for clients of officers trained in core correctional practices versus 49.9% for comparison clients — about 38% higher recidivism for untrained-officer caseloads in relative terms. That is a training comparison, not a direct high-versus-low fidelity contrast and not an AI effect. A later meta-analysis of 25 studies (Labrecque, Viglione, & Caudy, 2023) found favorable associations between higher implementation fidelity and rearrest and reconviction; the technical-violation result was not conclusive. The core correctional practices the rubric measures align with the skills named in the American Probation and Parole Association's national standards for community supervision (APPA, 2024). Together these findings motivate close attention to how supervision is delivered, without implying that automated grading itself reduces recidivism.

Review capacity is the practical constraint. If an agency reviews 1–2% of eligible visits today, covering the whole recorded corpus would be a 50- to 100-fold increase. That starting range is an illustrative operational scenario, not a measured national rate. Achieving full review would depend on recording coverage, staff capacity, and follow-through. Figure 1 separates the potential scale from the observed evidence.

3.2 What the first two papers established

Paper 1 evaluated AI grading against three independent expert graders across 30 visits spanning adult supervision, juvenile supervision, and treatment court (Green, Warren, & Meade, 2026). Its criterion compared the AI's distance from the pool mean of the expert graders with the median of the individual humans' distances from that mean; the AI passed a visit when its distance was no greater. The three-run Claude Sonnet 5 ensemble passed 23 of 30 visits (77%), and 16 of the 18 visits withheld from prompt adaptation. Mean absolute difference from the pool mean was 9.2 points and Pearson r was +0.778. The 77% describes visits passing that criterion, not item-level accuracy. Paper 1 showed performance within human scoring variability under that criterion; this study adds an adjudicated item-level reference.

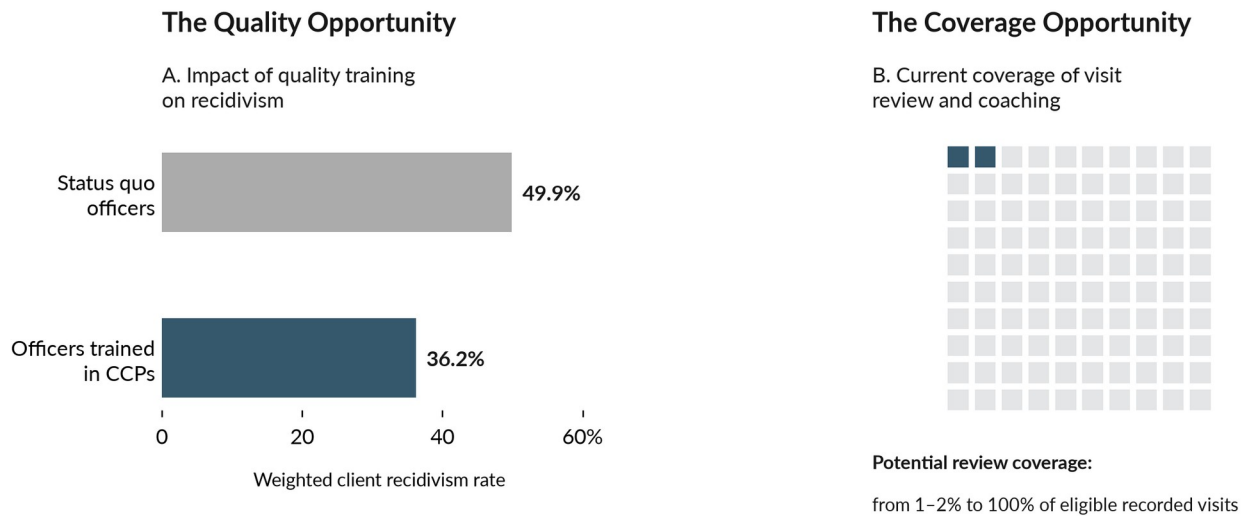


Figure 1. The quality opportunity and the coverage opportunity. Left: weighted client recidivism rates in Chadwick et al. (2015) for caseloads of officers trained in core correctional practices (CCPs) versus status quo supervision — a training comparison, not an AI effect. Right: an illustrative 2% review baseline; a 1–2% baseline implies a 50- to 100-fold increase if all eligible recorded visits could be reviewed. Neither baseline coverage nor expanded delivery was measured here.

Paper 1 also established that experienced professionals could assign substantially different scores to the same supervision visit. Pairwise human weighted-score differences averaged 19.0 to 33.3 points. Graders disagreed about whether an interaction met an item, whether an opportunity existed, and whether a partially completed skill deserved credit; the graders' Tier 3 NA counts averaged 3.8, 13.3, and 18.2 of 31 items. Paper 1 highlighted disagreement over role clarification and recommended that the graders work together to establish a shared scoring reference: “an adjudication study in which the graders confer to a consensus score ... that would furnish a tighter, agreed reference to measure the model against and turn the present agreement question into an accuracy one” (Paper 1, §7). This paper is that study, for ten selected visits.

Paper 2 evaluated 120 coaching outputs for the same 30-visit corpus — two human coaching sources and two LLM sources, assessed by one source-blinded expert reviewer across eight dimensions (Green & Warren, 2026). Claude Sonnet 4.6 had the highest mean composite rating, 3.58 of 5, against 3.30 for one human source, 3.30 for GPT-5.5, and 2.98 for the other human source; its advantage over the lower-rated human source survived multiplicity correction and its advantage over

the higher-rated source did not. Paper 2's coaching used the score context available at the time, not the reference-grounded grader developed here.

Together the first two papers established a plausible grading-and-coaching pathway while leaving the grading target unresolved. This study adds the reference-grounded grading layer recommended in both papers.

4 Methods

4.1 Design and corpus

This was a retrospective expert-adjudication study followed by adaptive model development, repeated testing of finalized workflows, and a cross-model extension. The corpus contained ten visits from Paper 1: five adult and five juvenile probation or parole visits, with no treatment-court visits. The ten visits were selected to examine human scoring disagreements and inconsistent AI grading, while also including comparison cases with greater agreement or higher-quality supervision. This was a selected, not random, sample; it does not establish representativeness across agencies, officers, populations, or visit types.

We used the de-identified transcript text with numbered speaker turns for evidence references. The original transcripts, rubric, initial human marks, adjudication workbook, and rationale records were preserved separately from model outputs. For readability the ten visits are labeled Visit 1 through Visit 10 throughout.

Following rubric-based evaluation work such as HealthBench (Arora et al., 2025), we evaluated the grader itself, examined performance by item group, and tested model and prompt dependence through repeated assessment. We borrow those methodological principles, described further in §6.4.

4.2 Expert adjudication and source reconciliation

Valerie Meade, Maureen Flanagan, and Katie Myers — the Paper 1 graders — formed the adjudication panel. Working from their original independent marks, the panel met together over a six-week period to review every item on which they had differed, discuss the transcript evidence, and arrive at a final adjudicated label with a written rationale. Where the original marks had already agreed, the agreed label was confirmed rather than re-litigated. We reconciled the extracted data against the original workbook cell by cell rather than treating blank extraction results as

missing judgments. The final record contained 192 YES, 193 NO, and 44 explicit NA decisions reached unanimously, plus one item on which the panel remained split.

The panel reached unanimous agreement on 429 of 430 rubric decisions. The single unresolved item concerned whether the officer identified and focused on the behavior being reinforced in one visit; two graders initially marked NO and one marked YES. The historical comparison file used the majority NO. Because the adjudication record does not show unanimity, we excluded this item from both model and reference denominators in the primary analysis and reported a 430-item sensitivity using the majority NO. The exclusion was based on the record, not on model correctness. The selected model matched the majority NO, giving 388/430 matches in the sensitivity analysis (Appendix A, Table 4). Appendix A retains the item identifier.

The panel's recorded discussion sharpened several distinctions that recur throughout the rubric: evidence that a skill step occurred versus the quality of its execution; independent credit for separate steps of a multi-step skill; an absent opportunity versus uncertainty; specific behavioral modeling versus general professionalism; and structured problem solving versus advice alone. The panel also settled the operational meaning of the cognitive-restructuring, reinforcement-balance, and motivational-interviewing items. These interpretations are part of the construct being evaluated, not incidental wording, and they are the direct answer to the NA-disagreement problem Paper 1 documented (§6.2).

Adjudication does not erase professional disagreement about what the rubric should reward. The record includes the concern that permissive interpretations could credit weak practice in a small corpus. We preserve that concern: a reference can make grading reproducible while the field continues to debate the rubric. The model-scoring instructions translated the recorded guidance into more explicit rules; those rules were developed against the reference and were not separately ratified as a new expert rubric.

4.3 Rubric and version boundaries

We retained the rubric's 43 items and original tier weights. The grading instructions clarified what evidence earns credit, how to score a partly completed skill, and when an item is not applicable (NA). These clarifications can change scores even when the rubric wording stays the same, so scores produced under different

instructions are not necessarily interchangeable. Appendix D summarizes the main clarifications.

The distinction between occurrence and quality matters most. An occurrence item asks whether a defined step is present, not whether it was skillful, well timed, or likely to change behavior. Quality remains a coaching question even when a step earns YES; conversely, a warm and competent visit does not earn credit for every specific technique. The distinction gives the grader a reproducible task and bounds what a score can be taken to mean.

4.4 Primary outcomes and development targets

The primary item outcome was exact agreement with the 429 unanimous reference labels. All three labels count; a model NA on a reference-applicable item is an error, not an exclusion. Cohen's kappa summarizes agreement after adjusting for the agreement expected from the two raters' label distributions (Cohen, 1960) — one means perfect agreement, zero means no better than the marginals predict. Appendix A also reports nominal two-rater Krippendorff's α (Krippendorff, 2011) and fixed-three-category Gwet's AC1 (Gwet, 2008) as alternative chance corrections.

Four development targets were set and recorded before the Astra search began: at least 90% exact agreement; overall kappa of at least 0.80; at least 85% agreement in every tier; and mean absolute error in the original-weight visit score of no more than five points. The 90% target exceeds the highest initial human agreement of 86.01%. These were AI development targets, not professional competency requirements; Table 1 displays the initial human results alongside them for comparison. Landis and Koch's (1977) "almost perfect" band begins at 0.81, distinct from the study's preset kappa target of 0.80.

Each visit's score was computed mechanically from the item marks, as in Paper 1. Each tier score is the number of YES marks divided by applicable YES-plus-NO marks; the visit score is 100 times the weighted sum with weights 0.20, 0.30, and 0.50. Model and reference scores use their own applicability decisions, so an NA disagreement can move a denominator as well as a numerator. An all-NA tier would be undefined rather than zero; none occurred in the primary comparison.

Mean absolute error (MAE) is the average absolute difference between model and reference visit scores; signed bias is the average difference. Because tier errors can cancel within a visit, we also report the weighted sum of absolute tier errors before

averaging across visits. Equal-tier and pooled-item score summaries are secondary sensitivities. Appendix A gives the definitions.

4.5 Adaptive development and autoresearch

Autoresearch here means an AI-assisted experimental loop for improving grading instructions and execution workflow (Figure 2) — the same guardrailed approach Paper 1 used, now aimed at an adjudicated target. A research agent examined current results and error patterns, proposed a general change, prepared a versioned candidate, and ran it against the development transcripts. A fixed evaluator compared the returned labels and mechanically computed scores with the reference. Those results determined whether a candidate was retained, revised, or rejected. We changed prompts and workflow organization, never model weights. The expert panel supplied the target; the research agent never created or revised a reference label.

We chose this approach because the central uncertainty was how to translate a judgment-intensive rubric into reliable instructions. The same model responds differently to a clearer opportunity rule, to independent credit for a skill step, or to splitting the rubric into separate grading calls. Comparing those alternatives fairly required many repeated executions scored identically. The agent-assisted loop made that practical and auditable: every configuration and its outputs were preserved, the same evaluator was applied, and regressions were visible alongside improvements.

All ten adjudicated visits were development data. The researcher could inspect reference disagreements and rationales, then propose general policy or architecture changes. Candidate grading calls received the transcript, rubric definitions, fixed instructions, and response schema — never the same-visit reference labels, human votes, correctness flags, or rationales. A small set of separately labeled review conditions, in which a second model call reviewed disputed items using other models' judgments with identities permuted, was kept apart from the standalone workflow and is not part of the selected result.

The Astra search moved from a historical-policy baseline through consolidated instructions, increasingly explicit criterion policies, whole-rubric grading, tier and skill decomposition, targeted review, and fixed repetitions: 269 grading calls across 13 adaptive conditions, then 60 calls for frozen repetitions of two policy families (Appendix C). Unsuccessful candidates were retained in the record. Policies and

repeat membership were frozen before the repetition phase; selection of the recommended family followed review of those results. We selected the whole-rubric criterion-v4 vote because it met all four targets with three calls per visit; a more complex two-policy vote matched one more decision but narrowly missed the five-point score-error target (§5.5).

Autoresearch produced a documented sequence of tests, including the finding that more reasoning, longer instructions, and deeper decomposition did not reliably help. Its controls also define its limits. The development process saw the same reference repeatedly even though individual grading calls did not. Archived prompts, preserved failures, fixed repeat membership, and a prohibition on retrying because agreement was low make the process auditable; they do not make it a held-out validation. That requires visits withheld from policy design.

4.6 Repetitions and voting

The selected workflow used three independently initiated, history-isolated whole-rubric calls per visit with the same fixed criterion-v4 policy and maximum reasoning setting. Independence here means separate executions, not statistical independence of their errors. For each item the majority label across the three complete outputs became the final mark. A three-way YES/NO/NA tie would require a separately recorded blind review and is never resolved from the answer key or response order; none occurred in the selected ensemble.

We measured pairwise agreement among the three runs and the number of unanimous item decisions, and we retained each complete run rather than whichever response happened to agree best on a visit. The final score was computed from the voted item labels, not averaged across three independently scored runs. That differs from Paper 1's three-run mean-score procedure and is a different endpoint.

4.7 Model transfer and model-specific development

To test whether the process transfers, we retained the completed GPT-6 Astra and GPT-5.6 Luna (OpenAI) workflows and ran three further models: Claude Sonnet 5 (Anthropic), Grok 4.6 (xAI), and Gemini 3.1 Pro Preview (Google). Luna received Astra's instructions unchanged with no model-specific optimization. For the other three we first tested the unchanged Astra policy, then ran a bounded development

pass over policy wording, reasoning settings where applicable, and whole-rubric versus tier-specialist execution.

For each additional model, selection ranked complete development candidates by primary 429-item agreement, with score MAE and simpler execution as secondary criteria; the baseline family could not select its best single repetition. Seven complete Sonnet candidates and six each for Grok and Gemini were evaluated. Each selected configuration was frozen before its own three fresh final repetitions, which reused the ten development visits and therefore measure post-selection performance on that corpus. Every selected final request was byte-identical to its selected development request; technical retries and blind tie calls were separately recorded.

Sonnet selected the original v4 policy divided into three tier-specific calls. Grok selected whole-rubric v4 with a clarification supplement. Gemini selected a compact operational guide with a clarification supplement, also divided by tier. Whole-rubric workflows needed three grading calls per voted visit; three-tier workflows needed nine (Appendix C, Table 6).

4.8 Validation and audit controls

Responses were checked for complete item coverage, permitted labels, required evidence, and valid transcript-turn references. Malformed or incomplete outputs were never converted into NO or NA. Technical retries repeated the frozen input and were triggered only by provider or validation failures, never by low reference agreement. Across the additional-model study, 655 of 670 attempts produced usable grades: 595 passed strict format checks and 60 needed only removal of an outer JSON code fence; raw responses were preserved and no label, citation, or internal structure was repaired. Fifteen attempts were unusable; Appendix C records the counts. Turn-reference checks establish syntactic validity, not independent semantic confirmation of every cited passage.

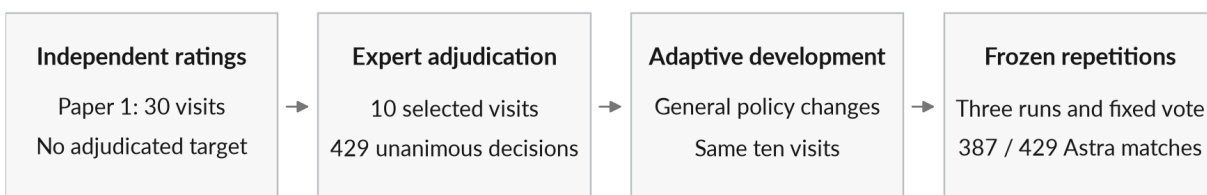
Grading sessions were isolated by instruction and audited for history and tool use. Model identifiers, reasoning settings, and access routes are recorded in Appendix C so that a future replication can hold them fixed; because vendors revise models frequently, a percentage reproduced with a different mixture of those components would not be a replication of the same workflow.

4.9 Analysis and uncertainty

We calculated descriptive counts, percentages, agreement statistics, score errors, class-specific errors, and tier, skill, and visit breakdowns. Human comparisons use each expert's original independent ratings against the later adjudicated reference. The three are labeled Human Grader 1–3; these labels do not correspond to Paper 1's grader numbering or to the author list. Human Average is the arithmetic mean of the three agreement rates – not a fourth grader, a majority vote, or the reference.

Comparison design. The initial independent ratings illustrate the consistency problem that can arise when review depends on individual expert judgment without a shared scoring interpretation. The experts graded before adjudication; AI development then used their resolutions on these same visits. This comparison evaluates the developed workflow against those initial ratings, not against experts newly trained on the final policy. The observed rates describe these experts and visits, rather than a field-wide benchmark.

The 429 decisions are nested in ten visits and were repeatedly exposed to development, so treating them as independent, prospectively sampled observations would overstate precision. A whole-visit bootstrap (ten visits drawn with replacement, 4,000 draws, seed 48271) gives a descriptive 2.5th–97.5th percentile range for pooled item agreement; it is a resampling range, not a confidence interval that corrects for adaptive selection. We make no significance claim that one model or the AI workflow is superior to future expert graders.



Additional model transfer and bounded optimization tested the portability of the workflow.

Development evidence throughout; no unseen-visit test in this study.

Figure 2. From independent ratings to reference-grounded grading. The ten visits were reused throughout development and final repetitions. Individual grading requests excluded reference labels; the research process used reference disagreements to refine general instructions. Fresh executions on these visits measure stability, not new-visit generalization.

5 Results

5.1 The panel reached near-complete agreement

The reconciled record supplied unanimous labels for 429 of 430 decisions (99.77%) – an adjudication-completion rate, not an estimate of independent inter-rater reliability. One item remained unresolved; the 430-item sensitivity analysis used the initial majority rating of NO. The primary reference is balanced: 192 YES, 193 NO, and 44 NA, so high agreement cannot be earned by predicting only the most common label. NA is 10.26% of the reference and occurs only within Tier 3, where the opportunity rules apply.

The experts' original independent marks agreed with their later reference at 86.01%, 65.50%, and 49.65% (mean 67.06%). All three revised some initial judgments through adjudication. The comparison design is described in §4.9.

5.2 The selected Astra workflow met every target

Astra's final fixed vote matched 387 of 429 labels (90.21%), with kappa 0.832 (Table 1; Figure 3). Original-weight visit-score MAE was 4.57 points and signed bias +0.99 points. All four development targets were met. Kappa 0.832 is "almost perfect" under Landis and Koch's (1977) convention; the alternative coefficients in Appendix A also indicate strong agreement on this corpus. Astra's agreement exceeded all three experts' initial ratings and was 23.15 percentage points above their mean.

Table 1. Preset AI development targets and observed results. Human columns show initial ratings against the later consensus. The tier row gives the lowest tier agreement; Met requires all three tiers to reach 85%.

Measure	Target	Astra	Human Grader 1	Human Grader 2	Human Grader 3
Exact item agreement	≥90%	90.21% Met	86.01%	65.50%	49.65%
Overall Cohen's κ	≥0.80	0.832 Met	0.760	0.470	0.245
Lowest tier agreement	≥85%	87.70% Met	80.00%	59.22%	46.00%
Weighted score MAE	≤5 points	4.57 Met	9.18	10.49	34.18

Using the same model and access route, agreement increased from 302/429 (70.40%) under the earlier grading policy to 387/429 (90.21%) – 85 more reference matches, or 19.81 percentage points – and score MAE fell from 14.06 to 4.57 points. That gain reflects the whole development process, policy changes and voting together.

The 430-item sensitivity with the majority NO on the split item gives 388/430 (90.23%). The two figures describe the same predictions under different treatment of one item; we lead with 429 because it does not assert consensus the record does not show.

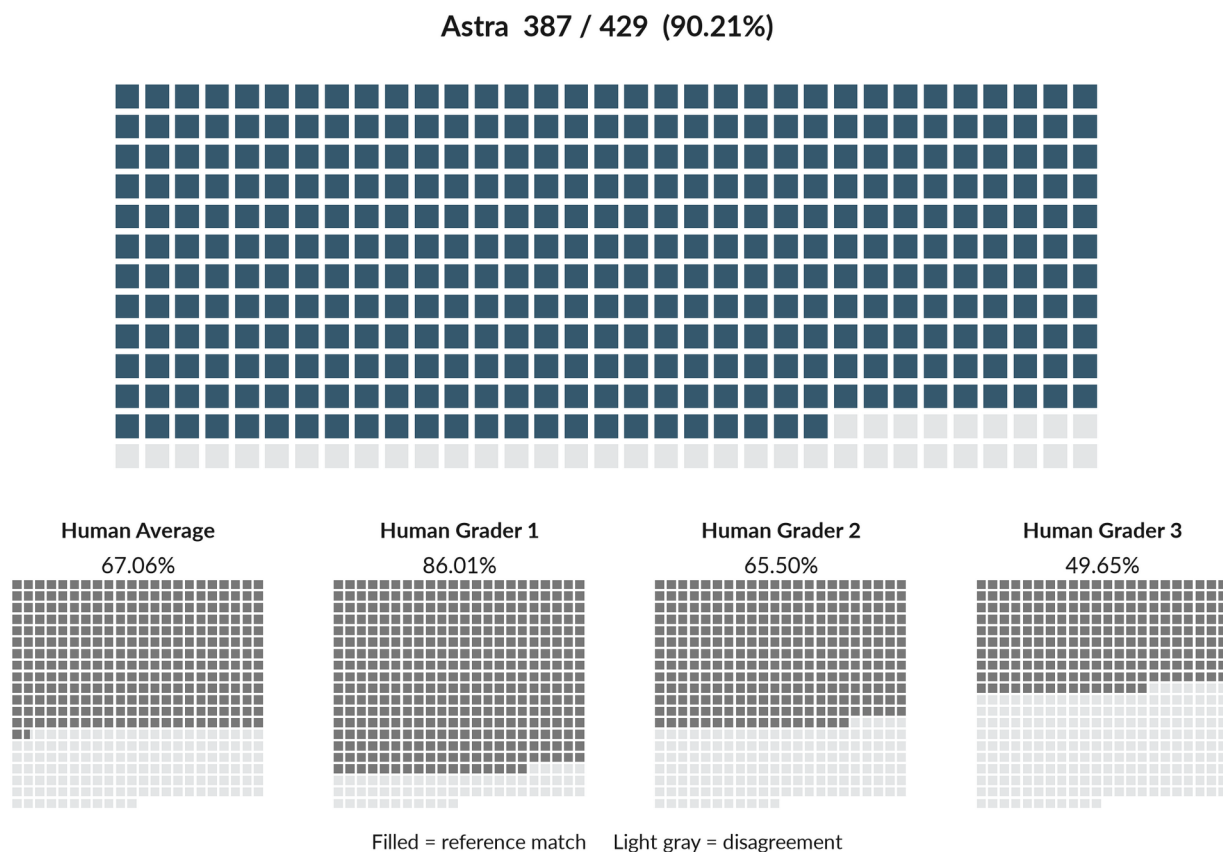


Figure 3. Agreement with reference — Astra 387/429 (90.21%). GPT-6 Astra and each human block contain 429 squares grouped by match status; filled squares are reference matches. Human Average shows the mean matching proportion and is not a fourth grader. Block sizes differ for layout only. The human ratings preceded adjudication and the clarified policy.

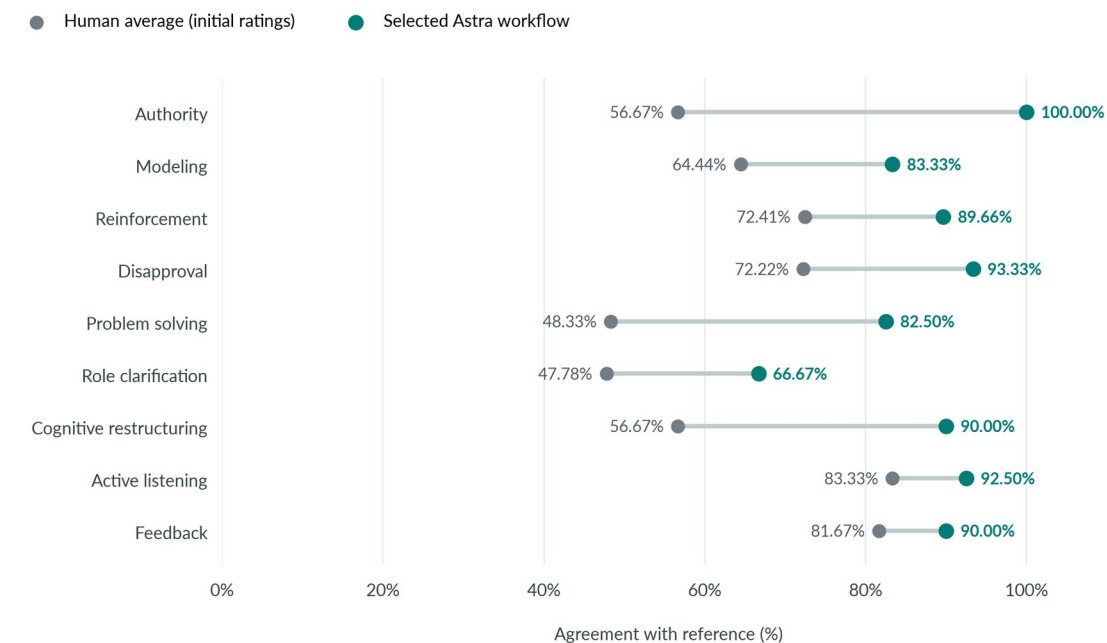
5.3 Astra exceeded the human average in every skill

Tier 1 agreement was 50/50 (100.00%), Tier 2 66/70 (94.29%), and Tier 3 271/309 (87.70%). All three clear the 85% tier target. Tier 3 holds 31 of the 43 items and

nearly all of the residual disagreement; its kappa was 0.797, just under the overall 0.80 target, which the protocol applied to the whole rubric rather than to each tier. The selected Astra workflow exceeded the mean initial human agreement in all nine Tier 3 skills (Figure 4), as well as in each of the three tiers. This comparison uses the same visits and expert-adjudicated reference. Modeling, problem solving, and role clarification had the lowest Astra agreement, at 83.33%, 82.50%, and 66.67%, respectively.

Astra exceeded the human average in all nine skills

Agreement with the expert-adjudicated reference across ten supervision visits



Human average is the mean of the three experts' initial independent agreement rates with their later consensus.
Both series use the same visit-item decisions within each skill domain.

Figure 4. Astra versus the human average across all nine Tier 3 skills. Points show exact agreement with the expert-adjudicated reference on the ten visits. Human Average is the mean of the three experts' initial independent agreement rates. Decisions per skill: Authority 30, Modeling 30, Reinforcement 29, Disapproval 30, Problem solving 40, Role clarification 30, Cognitive restructuring 40, Active listening 40, and Feedback 40. Reinforcement excludes the single unresolved item.

5.4 Error direction and score cancellation

The 42 primary disagreements comprised 21 false-credit YES marks on reference-NO items, ten missed-credit NO marks on reference-YES items, three NA marks on applicable items, and eight applicable marks where the reference was NA. The

model produced 205 YES marks against 192 in the reference — a modest residual tendency toward credit rather than a symmetric error pattern (Table 2).

Table 2. Selected Astra confusion matrix. Rows are reference labels; columns are model labels. Bold numbers show matching AI and expert-reference labels; other cells show disagreements.

Reference	Model YES	Model NO	Model NA	Total
YES	182	10	0	192
NO	21	169	3	193
NA	2	6	36	44

Class-specific recall was 94.79% for YES, 87.56% for NO, and 81.82% for NA; macro-F1 was 0.893. The lower NA recall matters because applicability moves both the label and the weighted-score denominator, which is why inspecting only total scores, or only the common labels, is not enough.

The 4.57-point average score error hides real variation across visits (Appendix B, Table 5). Five of ten visits were within five points and nine within ten. Visit 1 matched 33 of 43 items yet was over-scored by 16.13 points; Visit 5 had seven item disagreements and a total-score gap of only -0.45. Figure 5 shows why both views are needed: item agreement exposes disagreements a total can conceal, and score error shows what they could mean for an individual visit.

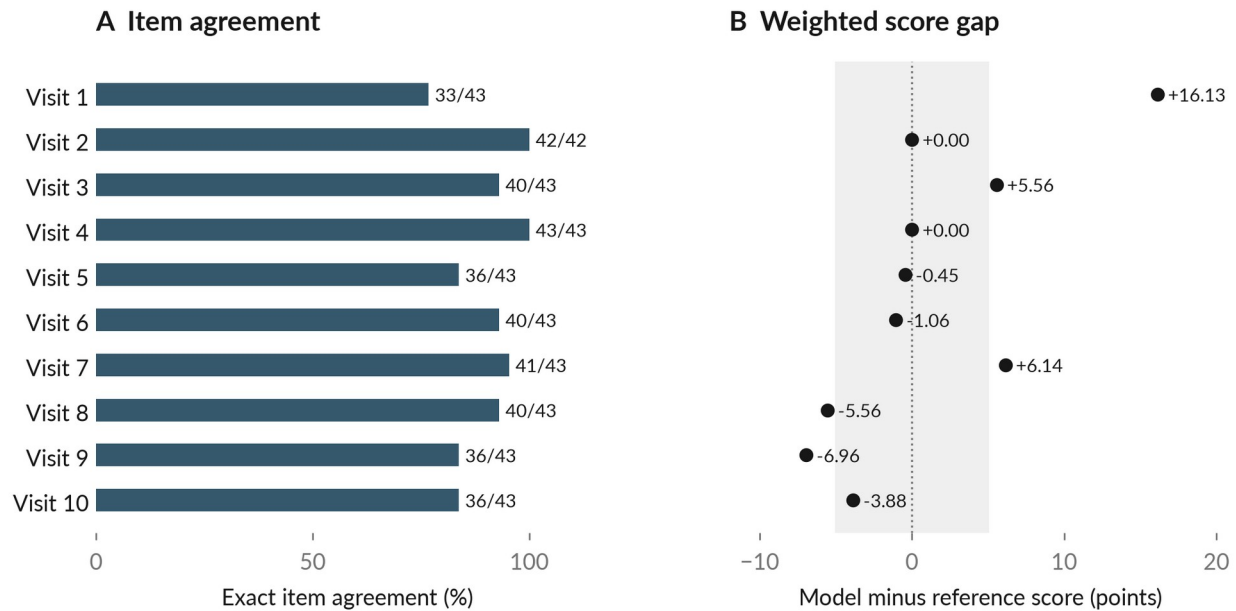


Figure 5. Selected Astra item agreement and weighted score gaps by visit. The dotted line marks zero gap; the shaded region spans ± 5 points. Positive values indicate model over-scoring. Visit 2 has 42 primary items; every other visit has 43.

Non-cancelling weighted tier MAE was 5.54 points; tier-specific score MAE was 0.00, 5.71, and 7.65 points; equal-tier MAE was 3.81 and pooled-item MAE 5.52. These summaries answer different questions. We report the original-weight result as primary and disclose the cancellation it can permit.

5.5 Repeatability and remaining disagreements

The three standalone v4 runs achieved 89.74%–89.98% agreement; none individually reached 90%, but the item-level vote did. Pairwise agreement among runs ranged from 96.04% to 98.37%, and all three agreed on 411 of 429 items (95.80%). Yet 35 of those unanimous items disagreed with the reference, and only seven of the 42 final errors fell among the 18 non-unanimous items.

Reviewing only items on which the runs disagreed would miss most disagreements with the reference. Repetition can reduce variability but does not challenge an interpretation shared across runs. Review sampling should therefore include unanimous items and the skills with lower reference agreement.

We report one selected three-member ensemble, not three independent successes. A different frozen two-policy vote matched 388/429 (90.44%) but had 5.04-point score MAE, narrowly failing that target; one individual run outside the selected family reached 389/429 (90.68%) but also missed the score-error target. For Astra,

Table 7 lists the selected whole-rubric v4 family, whose best individual run matched 386/429 (89.98%). We do not combine best values into a system that was never evaluated.

5.6 The process transferred; the models were not interchangeable

All five tested workflows exceeded the 70.40% baseline achieved by Astra using the earlier grading policy (Table 3). Astra alone met all four targets. Grok 4.6 reached 85.55% agreement with kappa 0.756; Luna, Sonnet 5, and Gemini reached 80.89%–82.05% (Figure 6). Under the Landis–Koch interpretation, Astra reached “almost perfect” agreement and the other four workflows reached “substantial” agreement. These results show why model, scoring policy, and execution workflow must be tested together rather than assuming a prompt will perform equally across models.

Table 3. Selected or retained final LLM workflows. All values use the 429-item primary reference.

Workflow	Matches	Agreement	κ	Score MAE	Targets met (of 4)
GPT-6 Astra (OpenAI)	387/429	90.21%	0.832	4.57	4/4
Grok 4.6 (xAI)	367/429	85.55%	0.756	8.40	0/4
GPT-5.6 Luna (OpenAI)	352/429	82.05%	0.688	6.75	0/4
Claude Sonnet 5 (Anthropic)	350/429	81.59%	0.692	7.01	0/4
Gemini 3.1 Pro Preview (Google)	347/429	80.89%	0.675	7.74	0/4

Of the three models given bounded model-specific development, Grok and Gemini improved on unchanged-policy transfer and Sonnet did not: Sonnet’s selected final vote fell from 354 to 350 matches even as its score error improved (Table 8). A configuration that wins a development pass does not necessarily win the later vote, which is one more reason to report fresh repetitions rather than the best development pass.

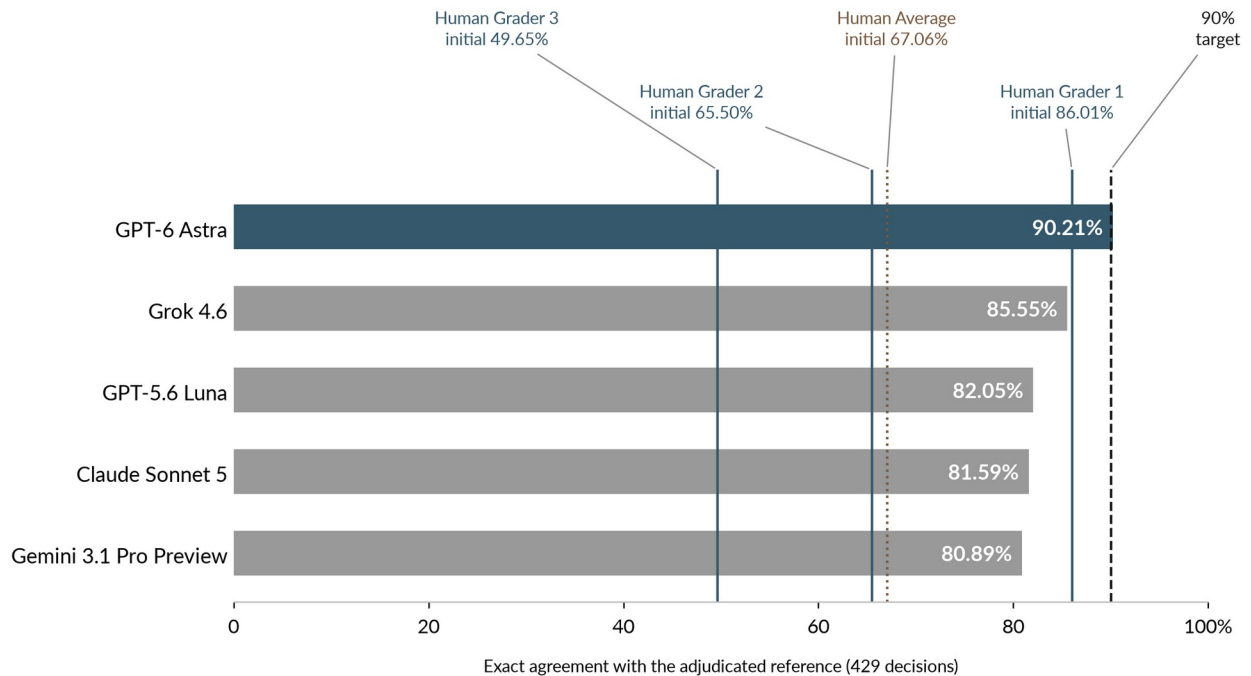


Figure 6. Five workflows, one process. Exact agreement of each selected or retained final workflow against the 429 primary decisions, with each expert grader's initial agreement and the Human Average as reference lines. Human Average is the mean of the three individual agreement rates. GPT-6 Astra clears the 90% target and all three initial human rates; every workflow exceeds the initial human average.

All three development peaks exceeded their final votes (Sonnet 364, Grok 376, Gemini 354 matches against final 350, 367, 347). The comparison reflects unequal development: Astra had the most prior development, Luna had none, and the other three had bounded searches with different selected architectures. The results rank the tested workflows under those conditions rather than the underlying models.

6 Discussion

6.1 Establishing and meeting the standard

The adjudication record shows that a nearly complete shared reference can be built for this corpus. The development result shows that a selected LLM workflow can meet it with strong exact agreement and small average score error. Whether every element of the reference has optimal construct validity, and whether the resulting score predicts outcomes, remain open questions.

The distinction between a rubric and its operational policy is central. General instructions leave expert graders with different reasonable readings. Explicit rules

make those readings visible and testable — and can also narrow what the instrument measures. A system that reliably detects the occurrence of a skill step may need a separate quality assessment to tell perfunctory use from skillful practice. The right response is versioned measurement, not a claim that one new score is the definitive measure of supervision quality.

The adjudicated labels and rationales are a research contribution independent of any model. Building on human consensus-coding methods (§6.4), they document how this panel resolved opportunity rules, applicability, and occurrence versus quality for the selected visits. Appendix D summarizes the grading rules developed from that guidance. The experts did not separately approve those rules.

6.2 What this closes from Paper 1

This study addresses several questions raised by Paper 1: establishing an adjudicated reference, clarifying NA decisions, recording evaluation targets, and examining repeatability. Role-clarification disagreements remain identifiable, and treatment-court cases remain outside this corpus.

An adjudicated reference. Paper 1’s abstract, §7, §8, and §9 all called for an adjudication study to supply an accuracy criterion. This paper delivers it for ten of the 30 visits, with 99.77% adjudication completion, a written rationale for every contested item, and a documented protocol that can be extended to further visits.

The NA problem. Paper 1 reported that the graders’ Tier 3 NA counts averaged 3.8, 13.3, and 18.2 of 31 — one grader marking nearly five times as many items inapplicable as another — and named “context provided opportunity” as the rubric’s weakest definition. Adjudication converged on 44 NA decisions (10.26%), all in Tier 3, all unanimous, with explicit opportunity rules for the disapproval, problem-solving, role-clarification, and feedback chains and a standing rule that cognitive restructuring is always applicable. The grader reproduced those NA decisions with 81.82% recall. The disagreement Paper 1 measured is now a written policy.

Role clarification. Paper 1 identified these items as the clearest NA split among expert graders. Here, they had the lowest Astra agreement (66.67%), although Astra exceeded the initial human average (47.78%). The recurring concentration helps identify where to examine disagreements; it does not by itself locate their cause in the rubric, the human interpretation, or the model.

A fixed analysis plan. Paper 1 set the expectation that the next study would fix its endpoint and metrics before grading. This study recorded four targets before the Astra search and reports them in Table 1 exactly as set. The recompute-from-item-marks policy Paper 1 adopted after finding arithmetic errors in the human score sheets was carried forward: every score here, human and model, is computed from item labels.

Run-to-run stability. Paper 1 reported two visits with large run-to-run swings in the grader's own scores; one of those visits is Visit 5 here. Paper 1's three runs scored it 37.2, 8.3, and 53.4. The selected workflow's three runs agreed pairwise on 96%–98% of items across the corpus, and Visit 5 landed within half a point of the reference. The selected workflow combined a fixed response schema, item-level voting, and explicit criterion rules; this study did not isolate their individual effects on stability.

Unscoreable visits and collateral contacts. Paper 1 asked for a resolution rule for visits a grader declines to score and a corpus-inclusion rule for collateral contacts. Those cases arose only in Paper 1's treatment-court visits, which were not in this adjudication subset, so the question remains open for the next tranche.

6.3 How strong the Astra result is

Ninety percent exact agreement across three label categories is a strong result on a judgment-intensive rubric. A kappa of 0.832 and the balanced reference distribution rule out a trivial majority-label explanation. Meeting the tier and score-error targets at the same time protects against an overall percentage that conceals a failed tier or a distorted total. Tier 1 was perfect; Authority was perfect; five of the nine Tier 3 skills were at or above 90%.

The evidence is strongest for feasibility on this corpus and weaker for generalization, because the same visits informed every stage of development. The whole-visit bootstrap range of 85.58%–94.65% describes corpus-level variability and does not correct for selection. The supportable claim is that the selected workflow met a demanding development standard set above the best human's initial performance; its accuracy on unseen visits remains to be measured.

The remaining disagreements also identify useful review priorities: role clarification had the lowest agreement, one visit was over-scored by 16.13 points, and most disagreements persisted across unanimous runs.

6.4 Relationship to established practice

Our approach resembles the methods OpenAI and Google used to evaluate automated graders in health care, drawing on expert-defined rubrics, human reference judgments, resolution of ambiguity, and direct testing of graders. OpenAI's HealthBench validated its rubric grader against physician judgments (Arora et al., 2025); HealthBench Professional added physician authorship, review, and adjudication and supports evaluation of ChatGPT for Clinicians (Soskin Hicks et al., 2026; OpenAI, 2026). Google evaluated its Personal Health Agent analysis-plan grader against individual experts and their majority judgment before using it for scaled assessment (Heydari et al., 2025).

Published results provide useful context. HealthBench adopted a GPT-4.1 grader with macro-F1 of 0.709 on binary rubric decisions. Google's Personal Health Agent analysis-plan grader achieved 82.4% agreement with individual experts and 86.9% with their majority, supporting its use for scaled assessment (Heydari et al., 2025). Our result of 90.21% agreement, macro-F1 of 0.893, and kappa of 0.832 demonstrates strong agreement with our expert-adjudicated reference using a related evaluation approach. These are methodological precedents, not a common leaderboard: tasks, labels, and reference procedures differ. Lyssn provides a related service-setting example: its automated counseling assessment supported feedback and supervision across 426 recorded sessions over nine months in an opioid treatment program (Peavy et al., 2024).

Community-supervision research also provides a precedent for consensus coding. The Strategic Training Initiative in Community Supervision (STICS) used independent trained coders who resolved differences to produce consensus ratings for analysis (Bonta et al., 2010). Although STICS included clinical feedback to officers, its consensus-coding process still required human review of the recordings. The present study extends this approach by testing whether an automated grader can reproduce an expert-adjudicated reference. It both evaluates grading agreement and contributes a means to expand review coverage with measured agreement with that reference, without requiring a panel to review each visit.

6.5 What the human comparison shows

On the same 429 decisions, the selected AI workflow matched the experts' consensus more often than any expert's initial independent rating: 4.20 percentage

points above the highest and 23.15 percentage points above the expert average. These initial ratings illustrate the consistency challenge of independent review on the study visits (§4.9).

The experts supplied the professional judgment underlying the reference, and AI development benefited from their disagreements and resolutions. Their work made a consistent evaluation target possible; the selected workflow reproduced that target with strong agreement, and expert judgment remains essential for interpretation and review.

6.6 Implications for coaching and review coverage

The pathway is reference-grounded grading, then specific coaching, then changed officer practice, then better participant outcomes. This study measures the first step. Paper 2 supplies separate evidence on coaching quality but used the score context available at the time, not this grader. Further study can build on both contributions to improve coaching (§8).

A broader review system makes specific examples available far more often than occasional audits can. Its value depends on useful feedback, manageable review burden, officer trust, and actual changes in practice. In the illustrative coverage scenario (§3.1), a 100-fold increase in graded visits is a change in scale; its benefit depends on what officers do with the feedback, and any reduction in rearrest must be measured directly. The opportunity is to make the feedback loop consistent and available, then measure whether it works.

7 Limitations

The corpus is small, selected, and fully exposed to development. These ten development visits do not establish how accurately the workflow will grade new visits, and repeated calls on the same material do not enlarge the sample. Adaptive selection makes the best observed workflow an optimistic estimate of performance on unseen visits even when its final requests are frozen.

The reference is expert-adjudicated, not independently verified truth. The panel's own earlier ratings were the material under discussion; one item stayed split; and the concern that permissive credit could reward weak practice was not eliminated by reaching consensus. The scoring-policy layer goes beyond the terse rubric wording and has not been ratified as a separate expert instrument.

The human comparison uses initial independent ratings, while the AI workflow benefited from subsequent reference-informed development (§4.9). It therefore does not estimate how experts would perform after calibration to the final policy.

Models did not receive equal development budgets, and their access routes and settings differed (Appendix C). Vendors revise models frequently, so a model name identifies a recorded configuration rather than a guarantee that a future call will behave identically. The study supports reproducible workflow descriptions within the preserved record, with limits on exact future replication.

The output audit verified completeness, label validity, and citation structure, not every semantic inference; high label agreement does not prove every accompanying explanation is faithful. And the study measured neither coaching quality with the new grader nor changes in officer conduct, participant engagement, or recidivism. The motivating literature supports the importance of supervision practice; it does not supply a causal bridge from this grading result to those outcomes.

8 Future work

In Paper 2, the best-performing AI coaching source received higher average quality ratings than either human coaching source in a blinded expert review of 30 visits. Further study, expert feedback, and advances in language models may improve coaching quality further.

Field evaluation follows the same chain: visit evidence, review prompt, officer response, support or referral, measured outcomes. A staged rollout can first examine coverage, review burden, coaching uptake, and demonstrated practice change; rearrest and other participant outcomes need longer follow-up and a design that handles baseline differences.

9 Responsible use and recommendations for agencies

This is a measurement study; we make three recommendations for agencies.

Keep the evidence visible. Officers and supervisors should be able to see the cited transcript evidence, understand the applicable policy, contest a mark, and distinguish a missed step from a judgment about overall competence. A score without its evidence is not a coaching tool.

Sample the unanimous items. Monitoring should include items on which the model's runs agreed, not only disputed ones, and should be stratified toward the skills with known weaknesses — modeling, problem solving, and role clarification in this version.

Re-validate on change. A model or policy change should trigger renewed evaluation against the reference rather than inherit validation automatically. Paper 1 said the same of prompt-model pairings; this paper's cross-model table shows why.

10 Conclusion

This study gives an affirmative and bounded answer to its question. Experts established a reference for 429 of 430 rubric decisions, and the selected workflow met it with 90.21% exact agreement, kappa 0.832, and 4.57-point mean absolute score error — strong agreement by every published convention we applied, on the corpus used to develop it. The additional models show the process transfers and that model, policy, and workflow must be tested together. A reproducible, reference-grounded grading layer is feasible. Reference-informed coaching is the next link in the chain.

Data and ethics

The transcripts represent real supervision visits used for this research under RePath's existing agency contracts, drawn from multiple community corrections agencies. All were de-identified before grading, with participant and officer identifiers removed, and the graders reviewed only de-identified text. The three expert graders were compensated for their time. The study did not assign participant interventions or make operational supervision decisions. The underlying transcripts contain sensitive justice-system information and are not released as a public dataset. Research materials may be made available by agreement subject to agency restrictions, privacy protection, and proprietary interests.

Competing interests and funding

RePath Labs designed, funded, and conducted this study and has a commercial interest in supervision technology, including grading and coaching applications. No external funding supported the work. Valerie Meade participated as a consultant and compensated expert grader; she did not compile the comparative scores or

operate the LLM experiments. The study recomputed scores from item-level marks, preserved unsuccessful conditions, and fixed candidate workflows before repeated evaluation. Independent follow-up remains important given RePath's commercial interest.

Author contributions

Ted Green led study conception, the research program, and manuscript development. Madeline Warren coordinated the adjudication study and contributed manuscript revisions. Valerie Meade contributed expertise to the rubric, the original grading, and the adjudication, and reviewed the manuscript. The named authors are responsible for the evidence, interpretations, and manuscript.

Acknowledgements

We thank consultants Maureen Flanagan and Katie Myers for their expert transcript grading, participation in adjudication, and detailed professional feedback.

AI tools assisted with prompt development, experiment execution, analysis support, figure preparation, and drafting. AI assistance is not expert adjudication and is not authorship.

References

- American Probation and Parole Association. (2024). National Standards for Community Supervision.
https://www.appa-net.org/eweb/docs/APPA/APPAs_National_Standards_for_Community_Corrections.pdf
- Arora, R. K., et al. (2025). HealthBench: Evaluating large language models towards improved human health. OpenAI technical report.
<https://doi.org/10.48550/arXiv.2505.08775>
- Bonta, J., Bourgon, G., Rugge, T., Scott, T.-L., Yessine, A. K., Gutierrez, L., & Li, J. (2010). The Strategic Training Initiative in Community Supervision: Risk-need-responsivity in the real world (User Report 2010-01). Public Safety Canada.
<https://www.publicsafety.gc.ca/cnt/rsrscs/pblctns/2010-01-rnr/index-en.aspx>

- 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>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- 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. https://doi.org/10.31219/osf.io/m7d5g_v1
- 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, version 1.0. https://doi.org/10.31219/osf.io/9wq32_v1
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29–48. <https://doi.org/10.1348/000711006X126600>
- Heydari, A. A., Gu, K., Srinivas, V., et al. (2025). The anatomy of a personal health agent. Google Research technical report. <https://doi.org/10.48550/arXiv.2508.20148>
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411–433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Krippendorff, K. (2011). Computing Krippendorff's alpha-reliability (literature updated September 2013). Annenberg School for Communication, University of Pennsylvania. <https://www.asc.upenn.edu/krippendorffs-alpha-reliability>
- Labrecque, R. M., Viglione, J., & Caudy, M. S. (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>

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>

OpenAI. (2026, April 22). Making ChatGPT better for clinicians. <https://openai.com/index/making-chatgpt-better-for-clinicians/>

Peavy, K. M., Klipsch, A., Soma, C. S., et al. (2024). Improving the quality of counseling and clinical supervision in opioid treatment programs: How can technology help? *Addiction Science & Clinical Practice*, 19, 8. <https://doi.org/10.1186/s13722-024-00435-z>

Soskin Hicks, R., Trofimov, M., Lim, D., et al. (2026). HealthBench Professional: Evaluating large language models on real clinician chats. OpenAI technical report. <https://doi.org/10.48550/arXiv.2604.27470>

Appendix A Measurement definitions

A1 Item labels and agreement

The primary dataset contains 429 visit-item pairs. Agreement for a pair is one if the predicted label exactly equals the reference label and zero otherwise; percentage agreement is 100 times the mean. YES, NO, and NA are all nominal categories. The historical sensitivity adds Visit 2, item 19 with its majority NO target, giving 430 pairs.

Cohen's kappa is observed agreement minus expected agreement, divided by one minus expected agreement, where expected agreement sums over labels the reference marginal proportion times the predicted marginal proportion.

Krippendorff's α uses nominal disagreement and finite-sample pooled marginals for the two raters. Gwet's AC1 uses the fixed three-category expected-agreement calculation. Their equal rounded values for kappa and α here reflect this dataset, not identity of the formulas. Undefined subgroup statistics are not coerced to zero.

For the selected Astra vote, nominal Krippendorff's alpha was 0.832 and AC1 was 0.862. An alpha of 0.832 exceeds Krippendorff's (2004) 0.800 guideline for drawing firm conclusions from coded data, with 0.667–0.800 reserved for tentative conclusions. He also ties the required level to the consequences of error, which is why we report class-specific errors and score MAE alongside the coefficients. AC1 uses a different expected-agreement model and also indicates strong agreement on

this corpus. These coefficients describe the same decisions under different chance models; their conventional thresholds are not interchangeable.

For a class, precision is the share of predicted instances matching the reference, recall is the share of reference instances correctly predicted, and F1 is their harmonic mean. Macro-F1 weights the three class F1 values equally. It is a secondary summary, not the selection criterion.

A2 Scores and applicability

For each tier, the score is YES divided by YES plus NO after the primary exclusion. NA is removed only from that rater's score denominator, not from item agreement. The visit score is 100 times $0.20 \times \text{Tier 1} + 0.30 \times \text{Tier 2} + 0.50 \times \text{Tier 3}$, preserving the historical weights while allowing model and reference applicability decisions to differ.

Total-score MAE averages the absolute model-minus-reference score gap across the ten visits; bias averages the signed gap. Non-cancelling tier MAE takes the absolute difference within each tier, applies the tier weights, multiplies by 100, and averages across visits. Equal-tier MAE gives each tier one-third weight. Pooled-item MAE combines all YES and NO items into one proportion before computing the gap. These answer related but distinct questions.

A3 Primary and historical sensitivity

Table 4. Primary (429) and historical (430) item-agreement sensitivity.

Workflow or grader	Primary matches	Primary rate	Historical matches	Historical rate
GPT-6 Astra	387/429	90.21%	388/430	90.23%
Grok 4.6	367/429	85.55%	368/430	85.58%
GPT-5.6 Luna	352/429	82.05%	352/430	81.86%
Claude Sonnet 5	350/429	81.59%	350/430	81.40%
Gemini 3.1 Pro Preview	347/429	80.89%	347/430	80.70%
Human Grader 1	369/429	86.01%	369/430	85.81%
Human Grader 2	281/429	65.50%	282/430	65.58%
Human Grader 3	213/429	49.65%	214/430	49.77%
Human Average	—	67.06%	—	67.05%

All target decisions use unrounded primary values. Human Average is computed from the three individual rates before rounding; its near-identical value under the two denominators is coincidental. The extra historical item is treated as a majority-rule NO, not as a newly unanimous decision.

Appendix B Detailed Astra results

B1 Per-visit agreement and weighted scores

Table 5. Per-visit results for the selected Astra vote.

Visit	Matches	Agreement	Reference score	Model score	Score gap
1	33/43	76.74%	39.37	55.50	+16.13
2	42/42	100.00%	59.08	59.08	+0.00
3	40/43	93.02%	19.69	25.25	+5.56
4	43/43	100.00%	33.12	33.12	+0.00
5	36/43	83.72%	43.73	43.27	-0.45
6	40/43	93.02%	52.01	50.95	-1.06
7	41/43	95.35%	50.50	56.63	+6.14
8	40/43	93.02%	65.21	59.65	-5.56
9	36/43	83.72%	75.03	68.07	-6.96
10	36/43	83.72%	88.31	84.42	-3.88

Reference and model totals use their own applicable-item denominators, with the split item excluded from both. Gaps are calculated before rounding. A positive gap indicates model over-scoring. Item agreement and score error diverge because the rubric weights tiers unequally, different item errors can offset, and NA decisions change denominators.

B2 Repeatability and alternative selections

The selected whole-v4 vote combined three complete fixed-policy runs with individual agreement 89.74%–89.98%; the vote reached 90.21%. Pairwise repeat agreement was 96.04%–98.37%. Unanimity on 411 items comprised 376 unanimous reference matches and 35 unanimous disagreements; among 18 non-unanimous items the vote matched the reference on 11 and disagreed on seven.

The two-policy vote reached 90.44% but exceeded the score-MAE target at 5.04 points and required six calls per visit. An exploratory, retrospectively fitted module

composite reached 91.14%; it was not eligible as a primary result. Its module-selection-only leave-one-visit-out sensitivity was 87.41%, which still does not remove the effect of tuning on all ten visits. Neither exploratory analysis replaces the selected workflow's result.

Appendix C Model workflows and execution

C1 Selected configurations

Table 6. Selected and retained execution recipes. Calls are scheduled grading calls per voted visit before technical retries or tie review.

Workflow	Policy	Grading unit	Calls	Access route (as recorded)
GPT-6 Astra (OpenAI)	Criterion v4	Whole rubric	3	OpenAI platform
GPT-5.6 Luna (OpenAI)	Unchanged Astra v4	Whole rubric	3	OpenAI platform
Claude Sonnet 5 (Anthropic)	Criterion v4	Three tiers	9	OpenRouter / Amazon Bedrock
Grok 4.6 (xAI)	v4 plus clarification	Whole rubric	3	OpenRouter / xAI
Gemini 3.1 Pro Preview (Google)	Compact guide plus clarification	Three tiers	9	OpenRouter / Google

Recorded model identifiers: GPT-6 Astra and GPT-5.6 Luna (OpenAI platform, maximum reasoning setting); anthropic/claude-sonnet-5, x-ai/grok-4.6, and google/gemini-3.1-pro-preview (API). Selected API requests allowed up to 65,536 output tokens for Sonnet and 32,768 for Grok and Gemini; Sonnet and Grok used the xhigh reasoning setting and Gemini high. Immutable model snapshots were not available for the platform-accessed models. These are recorded study settings; similarly named controls are not assumed to be computationally equivalent across vendors, and because models change quickly, a replication should record its own identifiers alongside the results.

C2 Execution counts and technical exclusions

The additional-model study comprised 670 grading attempts in total: unchanged-policy transfer baselines, development passes, and final repetitions. Development involved 111 Sonnet, 120 Grok, and 120 Gemini grading units (115, 121, and 123 attempts after technical retries); a tier-specialist pass required three units per visit.

Incomplete candidates were ineligible for selection, and unlaunched units were not counted as completed work. The final phase comprised 210 scheduled requests: 90 Sonnet, 30 Grok, and 90 Gemini. The remaining attempts were the unchanged-policy transfer baselines and technical pilots recorded in the research package. Of all 670 attempts, 655 produced usable grades and 15 did not.

Sonnet’s maximum-effort pilots returned no grades at output allowances up to 128,000 tokens; the usable baseline used the xhigh setting, selected on technical usability before accuracy review. The failure concerns that configuration, not the model.

Sonnet’s four three-way item ties in the final phase were resolved through two separately recorded blind tier calls. Grok and Gemini had no final three-way ties. The unchanged-policy baseline had one Sonnet tie and two Gemini ties, also handled separately. No incomplete ensemble was scored as complete.

C3 Individual final repetitions

Table 7. Exact primary agreement for each final repetition and its fixed item vote.

Workflow	Run 1	Run 2	Run 3	Final vote
GPT-6 Astra	89.98%	89.74%	89.98%	90.21%
GPT-5.6 Luna	82.05%	81.59%	80.42%	82.05%
Claude Sonnet 5	78.09%	81.35%	81.59%	81.59%
Grok 4.6	84.85%	85.55%	83.22%	85.55%
Gemini 3.1 Pro Preview	79.25%	78.55%	79.72%	80.89%

The final vote uses the three complete repetitions plus the blind tie procedure where necessary. A vote need not beat the best single repetition, and a single favorable run is never substituted for the ensemble. Every percentage uses the 429 unanimous reference items. Runs outside the selected family, including the single 389/429 run noted in §5.5, are not shown.

C4 Auditability and reproducing the numbers

The Astra archive contains 329 audited grading sessions, 469 prepared input packets (140 belonging to preparation-only alternatives that were not launched), and 27 immutable source files. The additional-model audit reconciled 670 attempts and verified 140 result assemblies including all selected final outputs; request identity checks covered all 210 scheduled final requests. Tests covered scoring

rules, primary-versus-historical exclusion, voting, technical-failure handling, and source preservation.

A replication should preserve the rubric text, operational-policy version, transcript preparation, item order, response schema, model identifiers and settings, repetition count, vote and tie rules, and mechanical scoring code, and should report failures and every selected repetition.

C5 Transfer and final workflow comparison

Table 8. Unchanged-policy transfer versus selected final vote. Change is in percentage points; score MAE in score points.

Model	Transfer	Final	Change	Score MAE
Claude Sonnet 5	82.52%	81.59%	-0.93	8.54 → 7.01
Grok 4.6	83.92%	85.55%	+1.63	9.69 → 8.40
Gemini 3.1 Pro Preview	75.99%	80.89%	+4.90	9.55 → 7.74

Appendix D Operational policy boundaries

The selected criterion-v4 policy asked the grader to weigh supporting evidence and counter-evidence, then decide whether the specific behavior was demonstrated in context. It credited a recognizable attempt at an individual step without requiring the whole multi-step skill, and it forbade assembling unrelated fragments into an intervention that never occurred. These are implementation rules developed against the reference, not separately ratified expert instructions.

Within Tier 1, the policy distinguished a purpose for the overall encounter from the purpose of a local administrative task, and a participant-owned goal from a court condition; redacted scheduling details could still establish a next-contact arrangement. Within Tier 2, structured cognitive-behavioral work was distinguished from advice or referral logistics; reinforcement balance was read across the encounter rather than as a literal four-to-one count; motivational interviewing required change-oriented engagement, not an isolated open question.

Within Tier 3, modeling required a recognizable behavioral example rather than routine professionalism. Role clarification required explanation of the supervision framework or responsibilities, not simply a command. Problem-solving steps could earn independent credit, but discussing a problem or listing services was not guided problem solving. Cognitive restructuring was always applicable. The

feedback chain required a distinct evaluative-feedback opportunity; absent one, the chain was NA rather than inferred from ordinary remarks.

These boundaries determine what a score means. Imperfect execution can earn occurrence credit, but the essential action must be present, and a missed opportunity is never treated as no opportunity. Whether these choices give the most useful measure of supervision quality is a construct-validation question beyond agreement with the present reference.