

Decision Models in a System Built for Language Models

How a new class of bounded-output AI fits an expert-referenced grading system for community corrections supervision

Ted Green and Madeline Warren (RePath Labs)

RePath Labs, eHawk, Inc.

ORCID iDs: Ted Green 0009-0008-9264-6568; Madeline Warren 0009-0000-0141-8810

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

RePath Labs working paper · Version 1.0 · October 15, 2026 · Not peer reviewed

Companion papers: Green, Warren, & Meade (2026a), DOI 10.31219/osf.io/9wq32_v1; Green & Warren (2026), DOI 10.31219/osf.io/m7d5g_v1; Green, Warren, & Meade (2026b), DOI 10.31219/osf.io/xk5u9_v1.

Suggested citation: Green, T., & Warren, M. (2026). Decision models in a system built for language models: How a new class of bounded-output AI fits an expert-referenced grading system for community corrections supervision. RePath Labs working paper, version 1.0.

© 2026 RePath Labs. All rights reserved.

Contents

Abstract.....	2
1 Introduction.....	3
2 Decision models as a new primitive.....	4
3 Methods.....	6
4 Results.....	8
5 Discussion.....	12
6 Limitations.....	14
7 Recommendations.....	14
8 Conclusion.....	15
Materials availability.....	15
Data and ethics.....	15
Competing interests and funding.....	15
Author contributions.....	15
Acknowledgements.....	16
AI assistance.....	16
References.....	16
Appendix A Development screens.....	18
Appendix B Measurement definitions.....	18

Abstract

A new class of AI models, often called decision models, answers each question by choosing from a fixed set of allowed answers rather than writing open-ended text. They are fast and inexpensive, and they cannot return content outside the answers they are given. Less than four weeks after the first of them, TypeSafe’s Jev, was released, one independent leaderboard was tracking more than 170, including models from Microsoft, OpenAI, Amazon and Cloudflare. We asked how well these models fit into an existing grading system for community corrections supervision visits, one built with expert human effort and developed around frontier large language models (LLMs). Using the expert reference standard from our previous study (10 visits, 43 rubric items, 429 unanimous YES/NO/NA judgments), we tested three decision models: TypeSafe Jev, OpenAI Decisions, and Microsoft Decision-1. Dropped into the system unchanged, they matched 65–77% of expert judgments. Adapting the task to each model raised agreement to 70–83%. The best decision-model workflows fell in the same range as three of five frontier LLM workflows, but below the best, GPT-6 Astra (90.2%). Every response was a valid label. Service calls took about a second or less and cost under half a cent per visit, against an estimated US\$2.22 per visit for the leading LLM workflow at retail prices. One configuration returned identical answers across three runs, disagreements with the experts included. Decision models are a useful new building block for community corrections technology, but they do not outscore LLMs on a system designed for LLMs. Systems should be built around the strengths of the tools they use and tested against a human reference to choose among them. The next step is a scoring design built for decision models.

Key terms used in this paper

Term	Meaning
Decision model	A model that returns one choice from a fixed answer set for each question (here YES, NO or NA). The name describes this answer format, not supervision decisions. In this study the model answers narrow rubric questions about an officer’s visit; people make all decisions about the person under supervision.
Frontier LLM workflow	A general-purpose large language model used with a developed grading policy, usually with several calls per visit combined by vote. These produced the Paper 3 results.
Reference standard	The 429 unanimous judgments reached by three expert graders in Paper 3. Agreement with it is the primary outcome.
Transfer	The existing rubric and instructions given to a new model unchanged.
Adapted condition	The same rubric and reference, with the task presented differently to suit the model. For OpenAI Decisions: P01 repeats the decision policy beside each question; P02 replaces a free-text output instruction with one suited to typed answers; P03 combines both. For Microsoft: MC01 adds compact policy reminders at each question and the typed-output instruction; C0 is its unchanged control.
Score MAE	Mean absolute difference, in percentage points, between model and expert visit scores across the ten visits (tier weights 20/30/50).
Under-credit / over-credit	Under-credit: expert YES, model NO. Over-credit: expert NO or NA, model YES.

1 Introduction

Community corrections agencies supervise far more people than their supervisors can review. Our earlier papers addressed one part of that gap. Paper 1 tested an AI grader for officer supervision visits against expert auditors (Green, Warren, & Meade, 2026a). Paper 2 tested AI-generated officer coaching in a blinded expert review (Green & Warren, 2026). Paper 3 built an expert reference standard for the grading rubric and showed that a developed GPT-6 Astra workflow matched 387 of 429 expert judgments (90.21%; Cohen’s kappa 0.832; visit-score difference 4.57 points) (Green, Warren, & Meade, 2026b). Each of those systems was built with frontier LLMs.

Since then a different kind of model has become widely available. Decision models return a choice from a fixed answer set instead of free text. TypeSafe released Jev, which it describes as the first “System One” model, in early access on September 15, 2026 (Hawley, 2026; TypeSafe AI, 2026). The category grew remarkably fast. Twenty-six days later, the independent JevBench leaderboard listed more than 170 decision models, including 148 open-weight models, many of them fine-tuned variants of the same base models (Benchmark Heaven, 2026). Large technology companies released their own within the same few weeks, including Microsoft (Microsoft-Decision-1; Srivastava, 2026), OpenAI (the Decisions API; OpenAI, 2026a), Amazon (the open-source Strands Decider 2B; Brooker et al., 2026) and Cloudflare (the Clef models; Chen et al., 2026). Their appeal is practical. Calls are fast and cheap. Because the output is restricted to the allowed answers, the model cannot invent a quotation, a citation or a rationale. Some configurations also return the same answer every time they see the same input. For public-sector software, where cost, auditability and predictable behavior matter, those are attractive properties.

This paper asks a direct question: **how well do decision models fit into an existing system built with expert human effort and frontier LLMs?** That is the situation most agencies and vendors will face. They already have a rubric, a policy layer and an LLM workflow, and they want to know whether a cheaper, faster, more predictable model can take over some or all of the work. We answer it by giving three decision models the Paper 3 task, first unchanged and then adapted to each model, and scoring them against the same expert reference used for the frontier workflows.

The short answer is that they work well but are not the best available. Specifically:

- **Plugged in unchanged, they fell short.** The three decision models matched 65–77% of expert judgments, against 81–90% for the frontier workflows.
- **Fitting the task to the model helped substantially.** Changing how the task was presented, without changing the rubric or the answer key, added 26 to 30 matching judgments for OpenAI Decisions and Microsoft Decision-1, and further development lifted Jev from 76.9% to as high as 83.0%. The best decision-model workflows reached 80–83%, the same range as three of the five frontier workflows.
- **They did not reach the top.** The best LLM workflow (GPT-6 Astra) stayed 7 to 10 points ahead overall, and no decision-model condition passed our stronger internal screen for over-credit.
- **Their operating properties held up.** Every response was a valid label. Calls took about a second or less and cost under half a cent per visit. One configuration returned identical

answers across three runs, while another vendor’s route changed 17 of 429 answers on identical requests.

We draw four conclusions. Decision models are a useful new primitive for community corrections technology. They are not a drop-in replacement for LLMs on judgment-heavy tasks, and they should not be expected to outscore LLMs on a system that was designed around LLMs. Systems should instead be designed around the strengths of the tools they use. And because decision models are now part of the toolset, choosing among tools has become an evaluation problem that needs a human reference to answer. The natural next study is a scoring design built for what decision models do well, tested against the original rubric on new cases.

Agreement with expert judgment carries real weight in this field. In this study, a label out of agreement with expert judgment can mean telling an officer they missed a skill they used, or crediting one they did not, and coaching built on those labels can mislead officers and erode their trust in the feedback. The broader decisions these tools may eventually inform, such as responses to violations, sanctions and supervision levels, affect people’s liberty, work and families. The right comparison, however, is with current practice, which relies on human reviewers. Against the same reference, the three expert graders’ initial independent ratings averaged 67.1% agreement (Green, Warren, & Meade, 2026b); the decision models in this study averaged 76.8% once the task was adapted to them. One configuration, run three times, also returned identical answers every time, while the experts’ independent ratings differed enough that the reference had to be built through adjudication. The decision models fall short of the best LLM results yet still outperform the human average, although the comparison favors them: the experts’ ratings came before adjudication, and the models’ instructions were refined on these same visits.

2 Decision models as a new primitive

2.1 What we mean by a decision model

We use *decision model* for any model used in a bounded-output role: it receives an input and a set of questions and returns one of the allowed answers for each. The three models tested here reach that role in different ways. Jev (TypeSafe AI, 2026; recorded as jev-1.13.0) is a specialist classifier. OpenAI Decisions (recorded as gpt-6-luna) is a typed decision endpoint that was in public beta at the time of testing (OpenAI, 2026a, 2026b). Microsoft Decision-1 was reached through OpenRouter’s Azure route (recorded as microsoft/microsoft-decision-1-20261009) (OpenRouter, 2026; Srivastava, 2026). They share an interface, not an architecture, and our results describe the specific versions and routes we used. Despite the name, these models do not make supervision decisions in this study or in the uses we propose. They answer narrow questions, such as whether an officer reviewed the participant’s goals, and the officer remains responsible for every decision about the person under supervision.

How they work. An LLM writes its answer piece by piece and can produce any text, image or code. A decision model is given the question and the allowed answers in advance. It scores each allowed answer and returns the one it rates most likely, usually with a probability attached. It writes no prose. The probability is useful in its own right because it tells a system how sure the model is, so low-confidence answers can be sent to a person or a more capable model for review. In this study OpenAI Decisions returned a probability with every answer. When the P01

configuration was at least 80% sure of its answer, it agreed with the experts 90.6% of the time. Those confident answers covered fewer than half of the judgments (201 of 429), and 19 of them still disagreed with the experts.

2.2 Claimed strengths, and what this study can test

Speed and cost. We measured service-call latency and token-based inference cost for every decision-model run. For the frontier comparison we have an estimate rather than a matched measurement: the selected Astra workflow, which uses three whole-rubric calls per visit, would cost about US\$2.22 per visit at current retail API prices. That is roughly 450 to more than 3,000 times the per-visit inference cost of the decision-model workflows (Table 6). The figure should be read as an order of magnitude. Batch pricing, prompt caching, fewer calls per visit and smaller models can all reduce it substantially, and the price of a given level of LLM performance has historically fallen steeply, by between 9 and 900 times per year depending on the task (Cottier et al., 2025). We did not have matched timing for the frontier workflows.

Bounded output. A decision model cannot write a fabricated quote or explanation, because it can only return an allowed label. That removes one familiar failure of LLMs. It does not guarantee agreement: a model can still return a confident label that the experts would not choose. We measured three things separately: **validity** (whether every answer was an allowed label), **agreement** (whether it matched the expert reference) and **consistency** (whether the model gave the same answers when the same task was repeated).

Repeatability. We ran one OpenAI Decisions configuration three times and compared two identical Microsoft requests, so we can describe repeatability on those routes over a short window. We cannot establish it as a property of the model class. Table 1, at the end of this section, summarizes how decision models and frontier LLMs compare.

2.3 Early evidence on decision models

Independent evaluations began appearing within days of Jev’s release, mostly as preprints. Their findings agree with each other and with ours. Li et al. (2026) compared Jev with sixteen generative and reward-model judges against blinded human adjudication. Jev came close to GPT-6 when the verdict could be read directly from the text, at much lower cost, but fell behind on math, code and logic. Sending only Jev’s low-confidence cases to a reasoning model beat GPT-6 outright at 41% of its cost. Deußer et al. (2026) tested Jev on 37 public datasets and found it outperformed two open LLMs on most of them, though all three models weakened on rubric-based judgments. Wang and Gao (2026) found that Jev outscored the strongest of four generative LLM judges at flagging security problems in AI agent activity logs, with responses in about a second. Together these studies suggest that decision models can match or beat LLMs on bounded tasks and fall short where a judgment requires reasoning. All of them use public benchmarks, however. None tests decision models against an expert-adjudicated standard in a professional field, or asks what happens when they are placed into a working system built around frontier LLMs. This study addresses that gap for community corrections.

Table 1. Frontier LLMs and decision models compared

	Frontier LLM	Decision model
What it returns	Open-ended text, images or code	One answer from a fixed list, usually with a probability
Explains its answer	Yes, in writing, though the explanation may not reflect how the answer was produced	No; only the label and its probability
Can make up content	Yes (quotes, citations, reasons)	No, only allowed answers; it can still return one experts would not choose
Shows how sure it is	Not by default	Yes, through the probability score
Same answer every time	Can vary between runs; Paper 3 combined three runs by vote	Consistent by design: identical results from OpenAI Decisions; less than 4% variation from Microsoft Decision-1 (17 of 429 labels)
Speed per call	Widely variable; depends on many factors	About 0.5 to 1.1 seconds in this study
Inference cost per visit	About US\$2.22 at retail prices (Astra estimate)	US\$0.0007 to US\$0.0049
Agreement with experts (same reference)	80.9% to 90.2% (Paper 3)	70.4% to 83.0% after adapting the task
Best suited to	Holistic judgment; nuanced or ambiguous evidence	Bounded questions with concrete, observable evidence
Main risk	Plausible but unsupported reasoning; higher cost	Confident answers out of agreement with experts

2.4 Why a system built for LLMs is a demanding test

The grading system in this study was not designed with decision models in mind. Its rubric describes officer practice in professional terms, and its policy layer (the operational rules for what counts as evidence, partial performance and an opportunity to act) was developed iteratively with frontier LLM workflows on these same ten visits. Many items ask for holistic judgment: whether an officer's attempt at modeling or problem solving was actually carried out, or whether the situation gave the officer an opportunity to act at all. The incumbent LLM workflow also used three calls per visit combined by vote, after extensive development.

That makes this a demanding test for a newcomer, and deliberately so. An agency that already runs an LLM-based system would face the same comparison. A strong showing here would justify switching; a gap tells us what has to change for decision models to be used well.

3 Methods

3.1 Reference standard and rubric

We reused the ten supervision visits and expert reference from Paper 3: five adult and five juvenile visits drawn from multiple agencies, selected to include both difficult and higher-agreement cases. The rubric has 43 items in three tiers: Foundations (5 items, weight 20%), Structure (7 items, 30%) and Core Correctional Practice Execution (31 items, 50%). Each item

receives YES, NO or NA. NA is a substantive judgment that the criterion did not apply, not a missing value. Three expert graders reached unanimous decisions on 429 of the 430 visit-item pairs: 192 YES, 193 NO and 44 NA. The one unresolved pair is excluded from the primary analysis, as in Paper 3. No reference label, rubric definition or tier weight was changed during this study.

3.2 Decision-model conditions

Jev. Archived development work tested self-contained criteria, structured named fields, decision trees, context reduction and the order in which the answers were offered. The final workflows combined three runs per visit by vote. We report a single initial pass, the fresh-control vote, and the range across six answer-order votes.

OpenAI Decisions. After an unchanged transfer, we tested three adapted conditions: P01 placed the decision method and the relevant policy rules beside each question; P02 replaced an inherited free-text output instruction with one suited to typed answers; P03 combined both. P01 was then run twice more to test repeatability.

Microsoft Decision-1. After a saved baseline, we froze a paired comparison before running it: a fresh control (C0) using byte-identical requests to the baseline, and an adapted condition (MC01) with compact policy reminders at each question and the typed-output instruction. Visit order was randomized, five visits running C0 first and five MC01 first.

Individual requests never contained the reference labels. The development process did use reference disagreements to design later conditions, so all ten visits are development cases, not a held-out test set.

3.3 Frontier comparators

The frontier results are the retained Paper 3 workflows: GPT-6 Astra, Grok 4.6, GPT-5.6 Luna, Claude Sonnet 5 and Gemini 3.1 Pro Preview. They differ from the decision-model conditions in development effort, interface and number of calls per answer. The comparison is between workflows as they were actually run against a common reference, not a controlled test of model architecture.

3.4 Outcomes and analysis

The primary outcome is exact agreement with the 429 reference judgments. We describe answers that differ from the reference as disagreements, not errors. The reference records expert consensus on a human judgment, not an objective ground truth, and a model answer outside it is not necessarily wrong. Secondary outcomes are Cohen’s kappa, score MAE, signed score bias and the direction of each disagreement, especially under-credit, over-credit and inappropriate NA. We report these separately because they can move in different directions: a change can raise agreement while increasing over-credit or leaving the score difference unchanged.

Before running the adapted conditions, we fixed internal screens for advancing a workflow to confirmation testing. The screens covered agreement, score difference, bias, over-credit, misuse of NA and tier and cohort floors (Appendix A). They were internal development rules, not external preregistration or deployment criteria. Changes from baseline were summarized with 10,000 whole-visit bootstrap resamples, which keep each visit’s judgments together.

Because the 429 judgments are clustered within ten purpose-selected visits that informed development, all results are descriptive.

4 Results

4.1 Plugged in unchanged, decision models fell short

Given the existing rubric and instructions without changes, Jev matched 330 of 429 expert judgments (76.92%; score MAE 11.10), OpenAI Decisions 312 (72.73%; MAE 16.56) and Microsoft Decision-1 279 (65.03%; MAE 17.85). OpenAI Decisions and Microsoft Decision-1 both under-scored officers relative to the experts: the OpenAI Decisions original pass scored visits 15.87 points low on average, and Microsoft Decision-1 scored every visit below its reference. The frontier workflows ranged from 80.89% to 90.21%.

4.2 Fitting the task to the model closed part of the gap

Changing how the task was presented, while leaving the rubric, weights and reference untouched, improved every vendor’s agreement (Table 2).

Table 2. Agreement before and after adapting the task to each decision model

Model	Condition	Matches	Agreement	Score MAE
Jev	Initial single pass	330/429	76.92%	11.10
Jev	Developed vote, fresh control	345/429	80.42%	7.04
Jev	Developed vote, best of six orders	356/429	82.98%	7.55
OpenAI Decisions	Transfer	312/429	72.73%	16.56
OpenAI Decisions	P01 policy at each question	340/429	79.25%	8.67
OpenAI Decisions	P02 typed-output instruction	319/429	74.36%	14.50
OpenAI Decisions	P03 combined	342/429	79.72%	7.47
Microsoft Decision-1	Fresh control (C0)	276/429	64.34%	17.90
Microsoft Decision-1	MC01 compact policy	302/429	70.40%	18.06

Score MAE is in percentage points across ten visits. Jev’s six answer-order votes ranged from 345 to 356 matches; the best was identified after seeing results on the same visits, so the fresh control is the fairer single figure. Microsoft’s saved baseline (279/429; 65.03%; MAE 17.85) is omitted here because C0 is its contemporaneous control.

For OpenAI Decisions, P03 resolved 48 baseline disagreements and introduced 18 new ones, a net gain of 30 judgments (+6.99 points), and cut score difference from 16.56 to 7.47 points. The bootstrap range for the agreement gain was +2.57 to +10.77 points. P01 achieved almost the same result (+28 judgments), and the difference between P01 and P03 was within noise. Repeating the policy beside each question did most of the work.

For Microsoft Decision-1, MC01 added 26 matching judgments over its contemporaneous control (+6.06 points; bootstrap range +3.50 to +8.60). It cut inappropriate NA answers from 52 to 18 and eliminated all nine NA answers on items that never allow NA. Its score difference did not improve (18.06 against 17.90), because changing an NA to NO adds a failed item to the score

and because MC01 did not reduce under-credit. A change that helped one vendor’s interface did not carry over unchanged to another.

Jev reached its best results only after the most development. Its largest single gain came from restructuring the criteria into structured named fields (352/429; 82.05%), a hint of what a scoring design built for these models might look like.

4.3 In the frontier range, not at the top

Table 3. Decision-model and frontier LLM workflows against the same 429-judgment reference

Workflow	Model class	Execution	Agreement	Score MAE
GPT-6 Astra	Frontier LLM	Developed, 3-run vote	90.21%	4.57
Grok 4.6	Frontier LLM	Developed, 3-run vote	85.55%	8.40
Jev, best answer order	Decision model	Developed, 3-run vote	82.98%	7.55
GPT-5.6 Luna	Frontier LLM	Transfer, 3-run vote	82.05%	6.75
Claude Sonnet 5	Frontier LLM	Developed, tier vote	81.59%	7.01
Gemini 3.1 Pro Preview	Frontier LLM	Developed, tier vote	80.89%	7.74
Jev, fresh control	Decision model	Developed, 3-run vote	80.42%	7.04
OpenAI Decisions P03	Decision model	Single pass	79.72%	7.47
Microsoft Decision-1 MC01	Decision model	Single pass	70.40%	18.06
Expert graders, mean of three	Human	Initial independent ratings	67.06%	17.95

Shaded rows are decision models. The expert row averages the three graders’ initial independent ratings before adjudication (agreement 49.65–86.01%; score MAE 9.18–34.18) from Paper 3. Frontier rows are the retained Paper 3 workflows; they differ from the decision-model rows in development history, interface and calls per answer. GPT-5.6 Luna is a different product from gpt-6-luna used through OpenAI Decisions.

The best decision-model workflows landed among the frontier workflows. Jev’s developed votes (80.42–82.98%) overlapped Gemini 3.1 Pro (80.89%), Claude Sonnet 5 (81.59%) and GPT-5.6 Luna (82.05%), and the single-pass OpenAI Decisions P03 (79.72%) was close behind. On score difference, Jev’s fresh control (MAE 7.04) and P03 (7.47) were in line with those frontier workflows. Neither decision models nor LLMs dominated as a class.

The top of the range, however, belonged to LLMs. Astra exceeded P03 by 45 judgments (10.49 points) and the best Jev vote by 31 (7.23 points), with lower score difference (4.57 against 7.47 and 7.55). Grok 4.6 also stayed ahead. On a system tuned for frontier LLMs, the best frontier workflows remained the most faithful to the expert reference.

4.4 Where decision models were strong

Table 4. Agreement by tier and skill: best single-pass OpenAI Decisions, best Jev vote and GPT-6 Astra

Tier or skill	Judgments	Reference Y / N / NA	Decisions P03	Jev (best)	GPT-6 Astra
Foundations	50	36 / 14 / 0	86%	86%	100%
Structure	70	38 / 32 / 0	87%	84%	94%
Practice execution	309	118 / 147 / 44	77%	82%	88%
Authority	30	23 / 7 / 0	90%	97%	100%
Modeling	30	8 / 19 / 3	63%	67%	83%
Reinforcement	29	13 / 16 / 0	86%	76%	90%
Disapproval	30	8 / 16 / 6	70%	87%	93%
Problem solving	40	11 / 25 / 4	65%	65%	83%
Role clarification	30	13 / 14 / 3	63%	67%	67%
Cognitive restructuring	40	4 / 36 / 0	95%	95%	90%
Active listening	40	33 / 7 / 0	73%	90%	93%
Feedback	40	5 / 7 / 28	85%	93%	90%

Tier rows partition the reference; skill rows partition Practice execution only. Reinforcement has 29 judgments because of the one unresolved pair.

Performance was uneven, which is the most useful finding for system design. Decision models matched or beat Astra on several skills: Jev on feedback (37/40 against 36/40) and role clarification (tie), both decision models on cognitive restructuring (38/40 against 36/40), and Jev within one judgment on authority and active listening. At the item level, P01 had more matches than Astra on 4 of 43 items and the same number on 14. The largest gaps were in modeling, problem solving and, for OpenAI Decisions, active listening and disapproval. These are skills where the grader has to judge whether an officer's attempt was actually carried out, not only whether something was said.

These comparisons need care. Each item has only ten judgments, and looking across 43 items makes a few favorable results likely by chance. The cognitive restructuring result in particular rests on a reference with only four YES judgments out of 40, so answering NO every time would already score 36. We treat the pattern as a hypothesis about where decision models fit, not as proof.

4.5 The disagreements that matter

Table 5. Direction of disagreements, original and adapted conditions

Disagreement type	Decisions original	Decisions P03	Microsoft C0	Microsoft MC01
Under-credit (YES → NO)	70	42	76	76
Over-credit (NO or NA → YES)	16	19	18	19
Inappropriate NA (YES or NO → NA)	19	18	52	18
Other (NA → NO)	12	8	7	14
Total disagreements	117	87	153	127

Most of the decision models' disagreements came from withholding credit. Adapting the OpenAI Decisions task cut under-credit from 70 to 42 but raised over-credit from 16 to 19. Microsoft Decision-1's 76 under-credit disagreements did not change at all across its conditions, even as its NA handling improved. The leading LLM workflow, OpenAI's GPT-6 Astra, leaned the other way. It gave 205 YES labels against the experts' 192 and scored visits 0.99 points above them on average, while the adapted decision-model conditions, OpenAI Decisions P03 and Microsoft Decision-1 MC01, gave 167 and 134 YES labels and scored visits 5.8 and 18.1 points below. Among the decision models, Jev came closest to the experts' balance, with 189 YES labels. The direction matters for coaching: a strict grader understates what officers did well, and a generous one overstates it. Both vendors' adapted conditions met our modest definition of improvement. None met the stronger screen: P03 exceeded the limit on over-credit (19 against a maximum of 16) and retained unresolved interpretation concerns on three items, and MC01 passed five of eight criteria but failed on score difference and over-credit (Appendix A). No condition replaced the existing grader, and no threshold was relaxed after results were seen.

4.6 Valid every time; repeatable on one route

All 60 OpenAI Decisions requests and every Microsoft Decision-1 request returned complete, valid answers: an allowed label for every item on every visit, and a defined score for every visit. No output had to be repaired or re-requested.

The three P01 runs returned byte-identical responses: the same 429 labels, probabilities and usage each time. That includes the same 89 disagreements with the experts. Repeatability made the model predictable, not more in agreement, and only the human reference showed which consistent answers departed from expert judgment. We confirmed separate requests for each run but could not rule out reuse of responses inside the provider, because no provider request IDs were returned. Microsoft behaved differently: its byte-identical baseline and control requests changed 17 of 429 labels. Repeatability therefore has to be measured route by route rather than assumed for the model class.

4.7 Speed and cost

Table 6. Operating measurements for decision-model workflows

Workflow	Calls per visit	Median call time	Inference cost per visit
OpenAI Decisions, transfer	1	0.48 s	US\$0.0018
OpenAI Decisions, P01 / P03	1	0.57–0.65 s	US\$0.0049
Jev, developed vote	3	0.60–0.61 s	US\$0.0049
Microsoft Decision-1, C0	1	1.00 s	US\$0.0007
Microsoft Decision-1, MC01	1	1.06 s	US\$0.0009

Times are client-measured service calls including transport. Costs are usage-derived from provider-reported tokens and published prices, not invoices. Both exclude transcription, preparation, orchestration and human review. For comparison, the selected Astra workflow is estimated at about US\$2.22 per visit at retail API prices (Section 2.2); matched timing was not available.

All 60 OpenAI Decisions requests across six passes cost about 23 cents in total; all recorded Microsoft work cost about 2.5 cents. Adding policy text at each question raised input size, by 2.66 times for OpenAI Decisions and 39% for Microsoft Decision-1, but costs stayed under half a cent per visit. At these prices, inference cost is unlikely to be the limiting factor for agency use. The gap to the leading LLM workflow is large today, roughly 450 times or more at retail prices, though that gap can be narrowed with the cost-reduction strategies noted in Section 2.2 and is likely to shrink as LLM prices fall. The costs that will matter are review time and resolving disagreements, which this study did not measure.

5 Discussion

5.1 Useful, but not a drop-in replacement

Decision models did well on a difficult, judgment-heavy task. Their best workflows agreed with expert graders about four times in five, which put them among capable frontier LLM workflows, and they did so with valid output every time, at a fraction of a cent per visit. That is enough to take them seriously as a component of supervision technology.

They were not a drop-in replacement. Unchanged, they trailed the frontier workflows by 4 to 25 points, and their gains came from reshaping the task for each model. Even then they stayed behind the best LLM workflow, and none passed our stronger screen for over-credit. The obvious approach of swapping a cheaper model into an existing pipeline and expecting the same or better results did not work here.

5.2 Why the incumbent won on its own ground

The system tested here was built for LLMs. Its policy layer was refined with frontier workflows on these same visits, its items often ask for holistic judgment, and the leading workflow used three calls per visit after extensive development. Decision models received comparatively short adaptation and, in two cases, one call per visit. The study cannot separate how much of the gap reflects the model class and how much reflects fit between task and tool. It does show that fit matters a great deal: a presentation change alone moved OpenAI Decisions' agreement by 7 points, and Jev's best results came from restructuring the criteria.

5.3 Building around strengths

The results point toward systems designed around the strengths of each tool rather than built for one class and fed to another. Several roles are worth testing:

- **Bounded sub-questions.** Applicability and opportunity checks, the presence of required visit structure, and other questions with concrete, observable evidence. MC01's sharp drop in inappropriate NA suggests decision models can handle applicability well when it is posed explicitly.
- **A repeatable second reader.** A consistent decision model can flag where an LLM grader's answer differs, directing human review to the disagreements.
- **High-volume first passes.** Cheap, fast screening of large caseloads, with an LLM or a person handling the judgment-heavy items and flagged cases.
- **Administrative tasks.** Missing fields, inconsistent dates, document completeness and routing, where the answer is bounded and the cost of a disagreement is low.

A hybrid design that routes each item to the tool that handles it best could, in principle, exceed either tool alone at lower cost. Li et al. (2026) report an early example outside this domain: routing only Jev's low-confidence cases to a reasoning model outperformed GPT-6 at 41% of its cost. On some aspects of decision-making, a decision model may outperform an LLM at lower cost and with greater consistency. These are hypotheses. Our item-level results are too small to choose a routing rule, and any such design must be fixed in advance and tested on new cases.

5.4 The next study: a scoring design built for decision models

The rubric was written to describe human supervision practice, and its operational rules were refined with LLMs. The next study should start from the same professional standards and build a version suited to bounded models: each criterion broken into observable evidence questions, explicit opportunity rules, and defined handling of partial performance. Domain experts would confirm that each decomposed version keeps the meaning of the original standard. The original and decomposed versions would then be compared on new visits that played no part in their design, using decision models, LLMs and hybrids, and measuring cost, latency and staff time as well as agreement.

One guardrail is essential. Breaking a difficult criterion into clearer evidence questions is a legitimate translation. Dropping the parts that models find hard would change what is being measured and could raise a score while weakening the standard. The aim is a better fit between tool and task, with the professional standard unchanged.

5.5 Choosing tools is now an evaluation problem

With more than 170 decision models now tracked alongside frontier LLMs, the useful question for agencies and builders is no longer which model is best, but which tool fits which task. This study shows why that question needs a human reference to answer. Overall agreement hid offsetting changes in under-credit and over-credit; a better label rate did not always mean a better score; and identical answers across runs included identical disagreements. A documented expert reference, with separate checks for the disagreements that matter, is what made those differences visible.

6 Limitations

Development data. All ten visits were purpose-selected and informed development, including the design of the adapted conditions. Results may be optimistic and do not establish performance on new officers, agencies or populations.

Unequal comparisons. The frontier and decision-model workflows differ in development effort, interface, calls per answer and date. The best Jev result was selected from six conditions after seeing results. The comparison describes workflows as run, not the causal effect of model architecture.

Small subgroups. Each item has ten or fewer judgments, and some skills have very few YES or NA labels. Item and skill comparisons are exploratory.

The reference is an operational target. It records what three experts agreed under this rubric and policy layer. It is not an independent measure of supervision quality or outcomes.

Repeatability and operations. Repeatability was measured on one OpenAI Decisions configuration over a short window and on one Microsoft comparison. Cost and time cover service calls only. The frontier cost is an estimate at retail prices, not a matched measurement, and staff time and end-to-end savings were not measured.

Hallucination was not tested directly. Bounded output prevents free-text fabrication by design; it does not prevent disagreement with expert judgment. The products tested were in beta or newly released and may change.

7 Recommendations

- **Models are not interchangeable.** Replacing one model with another, whether an LLM with a decision model or one LLM with another, will not produce the same answer to every question. In this study, decision models dropped unchanged into a system built for LLMs agreed with the experts 65% to 77% of the time. In Paper 3, frontier LLMs given the same rubric ranged from 80.9% to 90.2%. As Paper 3 concluded, the model, the instructions and the workflow have to be evaluated together, and any change of model calls for re-testing by whoever builds the system.
- **Design for the tool.** Present the task in the way each model handles best, and consider assigning different parts of a task to different tools. Never change the standard to raise a score.
- **Track the disagreements that matter.** Report under-credit, over-credit and misuse of NA separately from overall agreement.
- **Consistency is not the same as agreement.** Review repeated and high-confidence answers as well as uncertain ones.
- **Keep decisions with officers.** Use model outputs to organize information and flag items for review. A YES/NO output that affects sanctions, liberty or services remains a consequential judgment, however fast or cheap.

8 Conclusion

Decision models are a new primitive for community corrections technology. Placed into a grading system built with expert effort and frontier LLMs, they performed well, reaching the range of several frontier workflows at very low cost with valid output every time, but they did not match the best LLM workflow and did not succeed without adaptation. The lesson is not that one class of model wins. Systems should be built around the strengths of the tools they use and tested against a human reference to choose the right tool for each task. Decision models now belong in that calculation. The next step is to design and test a scoring approach that plays to their strengths while preserving the professional standard.

Materials availability

The rubric, reference labels, model outputs, scoring calculations, development screens and model-version details summarized in this paper may be made available to interested parties by agreement, subject to agency restrictions, privacy protection and proprietary interests. The transcripts are not distributed.

Data and ethics

This study reuses the de-identified supervision-visit transcripts and expert reference from Paper 3. The transcripts represent real supervision visits used for research under RePath's existing agency contracts, drawn from multiple community corrections agencies, with participant and officer identifiers removed before grading. The study did not assign participant interventions or make operational supervision decisions. Requests to Microsoft Decision-1 used Azure-only routing, no fallback, data collection denied and request-level zero data retention.

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. RePath Labs has no financial relationship with TypeSafe AI, OpenAI, Microsoft or OpenRouter beyond paying for API usage. eHawk, Inc. is enrolled in an entry-level Anthropic partner program; Anthropic had no role in this study, and one Anthropic model (Claude Sonnet 5) appears among the retained Paper 3 comparators.

Author contributions

Ted Green conceived and conducted the study, performed the analysis and wrote the manuscript. Madeline Warren helped draft and edit the manuscript. The named authors are responsible for its content.

Acknowledgements

The authors thank Valerie Meade, Maureen Flanagan and Katie Myers for the expert grading and adjudication that produced the reference standard used here. Their prior work made this study possible; they did not take part in the experiments reported here and are not responsible for its conclusions.

AI assistance

AI tools assisted with experiment preparation, analysis and manuscript drafting. The expert reference predates this work and was not generated or revised by AI. The authors reviewed and take responsibility for the content.

References

- Benchmark Heaven. (2026). JevBench: Decision model benchmark and leaderboard (board v1.7.46, updated October 11, 2026). Retrieved October 12, 2026, from <https://benchmarkheaven.com/jev-models>
- Brooker, M., Chambers, M., & Nonato de Paula, F. (2026, October 1). Introducing Strands Decider 2B: A small, open source, decision model. Strands Agents. <https://strandsagents.com/blog/introducing-strands-decider/>
- Chen, M., Reneau, A., & Flansburg, K. (2026, October 1). Introducing Clef: Our open-source decision models, and new RL fine-tuning platform. Cloudflare. <https://blog.cloudflare.com/clef-decision-models/>
- Cottier, B., Snodin, B., Owen, D., & Adamczewski, T. (2025, March 12). LLM inference prices have fallen rapidly but unequally across tasks. Epoch AI. <https://epoch.ai/data-insights/llm-inference-price-trends>
- Deußner, T., Sparrenberg, L., & Sifa, R. (2026). Evaluating and benchmarking the System One model Jev (arXiv:2609.37647). arXiv. <https://arxiv.org/abs/2609.37647>
- 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. (2026a). 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
- Green, T., Warren, M., & Meade, V. (2026b). 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. https://doi.org/10.31219/osf.io/xk5u9_v1
- Hawley, M. (2026, September 16). ChatGPT co-creator launches TypeSafe AI with “Jev,” a model built for machines, not people. VKTR. <https://www.vktr.com/ai-platforms/chatgpt-cocreator-launches-typesafe-ai-with-jev/>

- Li, Y., Miao, Y., Krishnan, R., & Padman, R. (2026). JEV-as-a-Judge: Accept when confident, escalate when unsure (arXiv:2609.26550). arXiv. <https://arxiv.org/abs/2609.26550>
- OpenAI. (2026a). Decisions. API documentation, accessed October 8, 2026. <https://developers.openai.com/api/docs/guides/decisions>
- OpenAI. (2026b). Create a decision. API reference, accessed October 8, 2026. <https://developers.openai.com/api/reference/resources/decisions/methods/create>
- OpenRouter. (2026). Microsoft: Microsoft-Decision-1. Model page, accessed October 11, 2026. <https://openrouter.ai/microsoft/microsoft-decision-1>
- Srivastava, A. (2026, October 9). Introducing Microsoft-Decision-1, our model for fast decision-making. Microsoft. <https://commandline.microsoft.com/microsoft-decision-1-model-foundry/>
- TypeSafe AI. (2026). Models. Jev documentation, accessed October 12, 2026. <https://docs.typesafe.ai/models>
- Wang, Z., & Gao, Y. (2026). JEV as a judge for agent trace security: An empirical comparison with generative LLM judges (arXiv:2609.34862). arXiv. <https://arxiv.org/abs/2609.34862>

Appendix A Development screens

Both screens were fixed before the candidate conditions were run and were applied at full precision without relaxation. They governed internal development decisions only.

Table A1. Stronger screen for OpenAI Decisions candidates, with the P03 result

Check	Requirement	P03 result
Valid answers	All 430	Pass (430)
Judgments in agreement	At least 325/429	Pass (342)
Score MAE	At most 14.56 points	Pass (7.47)
Absolute signed bias	At most 15.87 points	Pass (5.76)
Reference NO → model YES	At most 16	Fail (19)
Reference NA → model YES	Zero	Pass (0)
NA on applicable judgments	At most 19	Pass (18)
NA on never-NA items 32–35	Zero	Pass
Tier floors	34/50; 56/70; 222/309	Pass (43; 61; 238)
Cohort floors	Adult 160/214; juvenile 152/215	Pass
Unresolved interpretation concern	Blocks advancement	Fail (items 16, 17, 28)

P01 failed the over-credit limits; P02 failed the agreement floor, the NA→YES limit and the NA limit. All three met the modest improvement definition: complete output, at least 313 matches and score MAE below the 16.56-point baseline.

Table A2. Frozen Microsoft MC01 screen

Criterion	Requirement	MC01 result
Output completeness	All 430 labels valid	Pass (430)
Defined scores	All ten visits	Pass (10)
Agreement	More agreements than both controls	Pass (302 vs 279 and 276)
Inappropriate NA	Fewer than both controls	Pass (18 vs 51 and 52)
Never-NA violations	Zero	Pass (0)
Score MAE	Lower than both controls	Fail (18.06 vs 17.85 and 17.90)
Reference NO → YES	No increase over either control	Fail (14 vs 12; equals C0)
Reference NA → YES	No increase over either control	Fail (5 vs 3 and 4)

Appendix B Measurement definitions

Agreement. Exact matches with the 429 unanimous reference labels. A model NA on a reference-applicable item counts as a disagreement.

Visit score. Within each tier, YES divided by YES plus NO; NA items are excluded from that tier’s denominator. The visit total applies weights of 20% (Foundations), 30% (Structure) and 50% (Practice execution). Model and reference scores each use their own NA decisions.

Score MAE and bias. The mean absolute and mean signed difference between model and reference visit scores across the ten visits, in percentage points. Negative bias means the model scored visits lower than the experts.

Cohen's kappa. Agreement adjusted for the agreement expected from the two label distributions. Values: OpenAI Decisions transfer 0.54, P01 0.65, P03 0.66; Microsoft baseline 0.43, C0 0.42, MC01 0.49; Astra 0.83.

Bootstrap ranges. 2.5th to 97.5th percentiles of 10,000 resamples of whole visits, which keeps each visit's judgments together. They describe this development corpus and are not confirmatory tests.