LLM JUDGES VERIFY PRESENCE, NOT ABSENCE: OMISSION BLINDNESS IN GENERATED CLINICAL NOTES, ITS MECHANISM, AND WHAT RECOVERS IT

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges score generated text at scale, and the benchmarks that measure them ask whether what a text says is correct, not whether what it should say is there. We measure that second capability where it is consequential and, unusually, constructible: ambient AI scribes writing clinical notes, where published human audits find omission the dominant error class. Public corpora cannot supply the answer key - every clinician reference note we audited is materially discrepant with its own transcript - so we release 500 single-error note pairs built from transcript-derived, audited fact sheets: 298 in which a named fact is certainly absent, graded by severity and by surviving trace, against 202 added-or-altered controls. Across an eight-design ablation, paired discrimination (0.5 is a coin flip) reads 0.79–0.94 on added or altered content against 0.50–0.63 on omissions, and on single notes no design flags notes with omissions reliably more often than perfect ones. Criterion scope, output format, voting, wording and automated prompt optimisation all move the operating point without creating usable detection, each with a bound on what we could have missed, and a larger reasoning budget lifts discrimination without closing the asymmetry. What recovers detection is restructuring the task - list the facts the source establishes, then verify each against the text as a closed presence check. Two methods reach that restructuring independently, and a control separates its parts, the fact list carrying about a third of the recovery and closed per-fact verdicts the rest. Held out, reading those verdicts as a rule rather than averaging them into a score is worth 2.7 times the detection at an identical false-alarm rate, and transfers across judge families near-unchanged, while no threshold set on constructed pairs survives real scribe output. Omissions whose restatement survives elsewhere defeat both. A physician author validated 70 items in a sitting covering six stages, three of them structurally blinded, and sided with the pipeline on all 10 notes where it and the best monolithic judge reach opposite conclusions; the severity rubric has since been graded blind by an independent clinician - not an author - agreeing to within a grade across the companion census's two clinician sittings. We release the benchmark, the prompts and the judgements.

1 INTRODUCTION

An LLM judge is a second model asked to read a source document and a generated text and report what is wrong with the text. Judges make evaluation affordable at the scale generative systems operate at, and the benchmarks that validate them in clinical text (Ben Abacha et al., 2025) share a shape: an item either is correct or carries a planted error, and the judge is scored on finding it. That measures one capability and not the other a judge is deployed to provide: whether something the source established is missing. An error that is present leaves text to point at; an omission leaves nothing.

Clinical documentation is where that difference becomes measurable. Ambient AI scribes draft notes from consultation audio at scale, and published human audits of their output find omission the dominant error class outright, at 54% to 86.3% of all errors (Biro et al., 2025a; Anderson et al., 2025; Kernberg et al., 2024). A companion census of three deployed products over 142 consultations (565 notes) finds 31.3% [27.0, 35.6] carrying at least one verified failure, 24.8% once the two classes a patient record would have prefilled are set aside, the largest group being allergy and medication

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

information missing or wrongly recorded, and two of the three documented physical examinations on telephone consultations (Fox et al., 2026). The quality layer these products are checked by is an LLM judge. The domain is doing methodological work here rather than being the subject: a consultation transcript is an unusually complete record of what a note ought to contain, so ground truth in which a named fact is certainly absent can be built.

The nearest prior work establishes the gap outside medicine. AbsenceBench (Fu et al., 2025) gives models an original and an edited document side by side, finds insertions caught at 86.2 to 99.5% F1 against a mean drop of 56.9% on omissions, names an architectural cause (an absence supplies no keys to attend to), and repairs it with placeholder markers at the removal sites - a repair unavailable to an evaluator, who would need to know where the hole is. Ours is the setting its limitations name as untested and expect to be harder: semantic rather than surface absence, no side-by-side original, judges rather than diff engines, and the medical domain.

Contributions. (i) A released benchmark of 500 note pairs, 298 of them omissions in which absence is certain and *graded*, by severity and by how much of the removed fact survives elsewhere in the note. (ii) The presence/absence asymmetry isolated by ablation rather than inferred from one judge design, replicated on a second judge model family, with the commission arm anchored externally on physician-labelled data. (iii) A null across five remedy families, each with a bound on what we could have missed, the optimiser searching under consultation-level splits. (iv) A compute-matched localisation of the mechanism: the fact list accounts for about a third of the recovery, closed per-fact verdicts for the rest. (v) A per-fact decision rule with a deployable operating point that transfers across judge families near-unchanged, and the measurement bounding it: no threshold set on constructed pairs survives real scribe output.

2 Related work

Absence detection in language models. AbsenceBench (Fu et al., 2025) is the direct prior, positioned above; concurrent work finds the same asymmetry in the reviewing role on non-clinical answer keys (Chen et al., 2026). **Judge reliability in clinical text.** MEDEC (Ben Abacha et al., 2025) measures judges on physician-labelled commission errors, the mirror class to omission, and anchors that arm here; concurrent work (DeLucia et al., 2026) independently finds clinical completeness judges near chance, with rubric and prompting gains that reflect threshold shifts rather than separation. Appendix I takes four further studies in turn, including one reporting the opposite direction on prevalence-sensitive F1 (Dahlberg et al., 2026; Hussain et al., 2026; Bergman et al., 2026a; Vachhani et al., 2026). **Scribe audits** establish the error classes and omission's dominance (Biro et al., 2025a; Anderson et al., 2025; Kernberg et al., 2024; Taylor et al., 2026); in a separate planted-error study (Biro et al., 2025b) reviewing physicians caught omissions at roughly the same rate as objective errors, so the asymmetry here looks characteristic of machine judges rather than of review. **Summarisation factuality** is the literature the fix belongs to: atomic-fact decomposition (Min et al., 2023), QA- and entailment-based metrics (Scialom et al., 2021; Laban et al., 2022), and per-fact coverage in the lineage of RAGAS, a widely used open-source evaluation library (Es et al., 2024). MED-OMIT (Schumacher et al., 2023) is the fix's direct ancestor, fact enumeration for omission measurement on the same public corpus as one of our strata; CARE (Bedi et al., 2026) builds calibrated flags for hallucination and omission together, arriving by design where Section 7 arrives by measurement: presence and absence need separate detectors. **The gap.** None of these lines of work simultaneously tests semantic omissions in free clinical notes, with no answer key available to the evaluator, with severity and surviving trace graded, and with benchmark construction reported separately from judge performance.

3 The benchmark: absence made certain and graded

Testing whether a judge can find an omission needs notes in which a named fact is certainly absent, and no public corpus we examined supplies one. The obvious route is to borrow the answer key, since public corpora pair consultations with clinician-written reference notes and scoring against the reference is the standard recipe (Papadopoulos Korfiatis et al., 2022; Yim et al., 2023). Treating a reference note as truth assumes it is true: audited against their own transcripts, instructed not to penalise valid paraphrase, all 53 usable PriMock57 notes (100%, [93.2, 100]) and all 45 usable

ACI-Bench notes (100%, [92.1, 100]) carry at least one material discrepancy, and 345 of PriMock's 1,030 discrepancies of all grades are critical. Judging a scribe against these references imports 7 to 11 material errors into the answer key itself.

If reference notes cannot be truth, the transcript has to be. For each consultation we build a fact sheet by blind extraction from the transcript, audit it with three model critics, and use it to repair and verify the reference note, which becomes the clean note every pair is built from, complete against the sheet rather than the whole transcript (17 of 98 fail the strictest residual check; Appendix A; Figure A.2 is the schematic).The extraction instrument was itself measured blind against consultations we had written ourselves, recovering 646 of 650 authored facts (99.4%, [98.8, 99.9]) on the 27 of 30 re-extractions its critic panel kept. Severity is not natively stable: two frontier models grading the same details agree at a Cohen's κ of 0.177 [0.04, 0.31]. A written rubric fixes most of it, so the benchmark ships the grades of two cross-family graders working from that rubric, at κ 0.662 [0.59, 0.73] over the corpus's 683 graded facts (the 613 in Figure A.2's caption are the evaluation split alone), taking the lower grade where they disagree. Both rubric runs grade more severely than either model does natively, which we disclose wherever a result is conditioned on severity; graded blind by an independent clinician, the rubric reproduces to within a grade with no settled direction, so no claim here rests on its grades leaning one way (Appendix A).

Removing a fact should then be mechanical, and it is not, because real notes repeat themselves: mapping every place each fact is stated in the clean note shows 574 of the 1,791 present facts (32.0%) stated more than once, and a cross-family panel verified only 71 of 126 complete-removal attempts (56.3%), so 44% failed, against 114 of 151 (75.5%) for partial removals that leave a trace by design; the 92 rejected pairs are excluded rather than patched (Appendix A itemises why). The 185 pairs those attempts produced are the newly built cohort, one of four sources the released set pools and Appendix A tabulates, with provenance a field on every pair. A benchmark built only from complete omissions therefore tests an unrepresentative slice: in real notes the complete case is the minority that survives construction.

What the released set holds. 500 pairs over 117 consultations, of which 495 over 112 consultations form the evaluation set. An **omission** pair removes a fact from the note and a **commission** pair adds or alters one. The evaluation set's omission half is 293 pairs: 150 complete removals, 86 where a **fragment trace** survives (a fragment of the removed statement or its immediate neighbour) and 57 where a **restatement trace** survives (a second, independently stated mention of the same fact), against 202 commissions and 112 clean twins, each the same note with the fact intact. The commission controls are carried over from an earlier build while most omission pairs are newly built, so the headline asymmetry compares across construction cohorts; the commission side is anchored externally (Section 5), and a robustness slice restricted to the uniformly built omission pairs *widens* it. The holdout has two layers. An earlier, exploratory optimiser worked on a 22-consultation development pool carved out before the released set was built; the confirmatory campaign (Section 6) then partitioned the 112 evaluation consultations, learning and accepting on 65 and holding out the remaining 47, whose 151 pairs were committed before that search began, never seen by any optimiser.

4 THE JUDGES UNDER TEST, AND HOW DETECTION IS SCORED

Everything that follows is scored two ways. **Paired discrimination** asks whether the judge scores the flawed note strictly below its own verified-clean twin: the share of pairs ordered correctly plus half of any ties, a paired win rate in the same family of measure as AUC, where 0.500 is a coin flip. It is a capability ceiling, because production never has the twin. **Single-note detection** is the deployment measure: one note arrives, the judge flags it or does not, and that flag rate means nothing except beside the same judge's false-alarm rate on clean notes. Each design carries a **flag rule**, the fixed threshold or verdict that turns its output into a flag, and "at its own rule" below means measured under that rule. Every remedy below is judged by whether it slides those two rates along a curve or lifts it; the two are never combined. Four sets of notes carry every judge number measured on this paper's own benchmark: an exploratory pass, the 151-pair held-out subset, the full evaluation set and the companion census's real vendor notes (Fox et al., 2026); no rate is compared across two of them except where the sentence says so, and Table K.1 maps which claims live on which set.

Five groups of judges share one pinned GPT-family judge model for every judging call, at matched settings (the pipeline's extraction, audit and second-look stages run on their own pins; roster Ta-

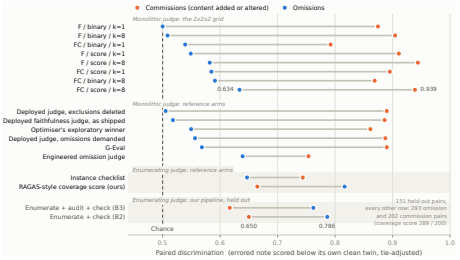


Figure 1: Paired discrimination on omissions against commissions, per judge design. Tie-adjusted probability that the errored note scores strictly below its own clean twin, plus half the tie mass; chance is 0.500. The sets read separately: monolithic and reference-judge rows are the 495 evaluation pairs (293 omissions, 202 commissions; eight designs at 3 replicates, reference judges at 2, the coverage-score row on 289 and 200 after 14 parse failures), the two pipeline rows the 151 held-out pairs (131 omissions, 20 commissions), means of 3. F/FC = faithfulness / faithfulness-plus-completeness criterion; binary/score = output format; k = samples. The RAGAS-style coverage score is the coverage component of Section 7's recipe.

ble K.2, pins Appendix D), including how much it may deliberate: every ablation design and reference judge runs at no reasoning effort inside a 1,024-token answer cap - chosen to reflect the cost and latency budgets an evaluation actually runs to in the field, inside a deployed application - and Section 7.2 measures what raising that budget does. A judge is **monolithic** when it reads the whole note with one prompt and returns one holistic judgement, against the enumerating designs, which list the transcript's facts first and check each one. The reference judges include a faithfulness judge deployed in the authors' own production evaluation - **the deployed faithfulness judge** below (Appendix H) - and **the RAGAS-style recipe** below, a coverage recipe in the RAGAS lineage, is our own implementation, its prompts disjoint from the pipeline's. The ablation crosses the three design choices underneath any deployed judge: **what the judge checks**, that everything in the note is supported by the transcript, or that plus whether anything important is missing; **how it answers**, a yes/no verdict or a 0-to-10 score; and **how many times it is asked**, once, or as an eight-call ensemble combined by a winsorised mean, the extreme scores replaced by the nearest kept values before averaging. That gives **eight designs** differing one decision at a time, deliberately more minimal than anything shipped. Each ran three times: 14,568 judgements over 65,556 model calls, zero recorded parse failures, and the same verdict on the same note across runs 93.2% of the time. Replicates collapse to a per-note majority before any significance test, and bootstrap resampling is at the consultation rather than the note, because pairs from one consultation share a transcript, a clean twin and a fact list; comparisons between two methods' flags are collapsed the same way before testing. Paired flag comparisons use an exact McNemar test with a Holm correction across families; proportions carry Wilson 95% intervals, agreement coefficients cluster-bootstrap ones; and where we claim something made no difference we state how large a difference we could have detected rather than a p-value above 0.05.

5 JUDGES VERIFY PRESENCE, NOT ABSENCE

Across all eight ablation designs, judges separate commissions from their clean twins and do not separate omissions. On added or altered content the eight run at 0.792 to 0.944 paired, six at 0.87 or above (Figure 1; every cell is Table K.3); on omissions the same eight read 0.500 to 0.634. Inside the best of the eight (faithfulness plus completeness, scored, eight samples - the best monolithic judge throughout this paper), what signal remains concentrates where it helps least. By how much of the fact survives, the signal falls from 0.690 on complete removals to a coin-flip 0.526 on restatement traces (Figure K.4). By severity: critical 0.683, supporting 0.586, peripheral 0.568 - a real, modest gradient, present only in the best design, and smaller than the trace gradient. The matched-fact contrast between complete and partial removals is $p = 0.022$ uncorrected and does not survive the family-wise correction Section 4 sets, so it is a supported pattern rather than a proven effect.

On a single note, nothing usable survives. Those are ceilings measured with the twin available. Among the scored designs the largest detection-to-false-alarm gap is 1.02 standard errors, which is noise, and two of the eight run negative. The two yes/no completeness designs flag far more notes - and a third of perfect ones - but their gap is 1.2 to 1.7 standard errors once replicates collapse to a per-note majority (Table K.3). Nor is the flag line the artefact: threshold-free, AUC runs 0.503 to 0.575 across the eight, and swept to a 10% ceiling the best reaches 8.3% detection at 6.5% false alarms. What is missing is separability, not calibration.

The same asymmetry in every judge family we ran. The reference judges reproduce it - the deployed faithfulness judge at 0.518 paired on omissions against 0.886 on commissions, the checklist judge (a transcript-generated checklist of roughly 25 items) at 0.647 against 0.744 (Table K.5) - except the RAGAS-style recipe, which inverts it (Section 7.5). The asymmetry also survives a change of judge model family: two of the eight designs re-run with a Gemini-family judge, prompts unchanged to the character, read commissions 0.951 and 0.961 against omissions 0.551 and 0.683, the

restatement-trace case at chance in every replicate of both. One difference is genuine: the second family's completeness-scored design clears its own noise on single notes in all three replicates (19.1 to 20.5% detection at 0.9 to 3.6% false alarms), which no primary-family design did. Single-note detection is therefore judge-dependent, and the single-note claims above are scoped to the family measured; the asymmetry and its conditioning replicate. That check carries one confound, the second family's endpoint refusing a no-reasoning setting (Appendix G).

The commission arm holds on physician-labelled data. MEDEC (Ben Abacha et al., 2025) is physician-labelled clinical text with no connection to this study, each item error-free or carrying exactly one substitution error, so it holds no omissions to detect. It also has no source document - the prompt's transcript slot carries a no-source placeholder - so what it anchors is spotting a wrong statement rather than transcript verification. Re-running the best of the eight on a 300-text stratified sample frozen before judging gives paired 0.827 and AUC 0.811; the matched contrast with the same design on our omissions is the AUC pair, 0.811 against 0.575, so the commission side is externally anchored. On those same labels the widest remedy of Section 6 behaves as it does here, an added omission clause raising detection and false alarms together while nothing separates better underneath (AUC 0.800 against 0.811). The omission arm has no external anchor, for the reason Section 8 gives.

6 Five remedies move the operating point; none restores usable detection

We tested five families of remedy: what the judge checks, how it answers, how many votes it gets, how the prompt is worded, and letting an optimiser rewrite the prompt. Every one shifted the operating point along the same weak curve without producing reliable single-note separation, which Figure K.2 draws as a cloud of designs on the no-signal diagonal - all five at the compute settings Section 4 records, the optimiser searching inside that same budget.

Scope, format and voting move the point, not the curve. Their effects are modest and additive (Table K.4), the widest span being criterion scope's $+0.036$ to $+0.082$ paired per design and the smallest voting's. Together they take paired discrimination on omissions from 0.500 to 0.634 while the deployment measure stays at noise. **Wording moves it once, through an instruction rather than a deletion.** The deployed faithfulness judge ships with the line "omissions are NOT errors", and we ran it in three wordings differing only in how they treat omissions: as shipped, with the exclusion deleted, and with an affirmative instruction to report omissions added. Deleting the exclusion does not help and if anything hurts. The affirmative instruction is worth $+0.051$ paired against the deleted arm and roughly doubles detection at an unchanged false-alarm rate - the same 27 of 224 clean-note judgements flagged under all three wordings, though not the same notes - the one place in the study where wording moved detection at no measured cost in false alarms (Table K.4).

Prompt optimisation does not survive a held-out test. We used reflective prompt evolution (GEPA) (Agrawal et al., 2026), which rewrites a judge prompt from traces of its errors and keeps rewrites that score better. We ran three campaigns and rest on the third, which learnt and accepted on the 65 consultations Section 3 names and touched the held-out 47 once, at the end (Appendix B). No gain survived. Aimed at the pipeline's check stage the winner gained 5.2 points of paired discrimination on the acceptance split and came back 2.3 points worse held out, on an interval excluding gains larger than about 2.7 points; aimed at the best monolithic judge it returned $+0.004$ paired, a fifteenth of the noise a single baseline run carries on these pairs (Table B.1). Even on its own acceptance set the first target's gain was not significant (8 wins against 3, $p = 0.23$) - a 20-consultation validator will hand a team an improvement a 47-consultation held-out set does not see.

The execution setting decided what the search appeared to find. All three campaigns executed their candidate prompts at the judge settings of Section 4 - no reasoning effort inside a 1,024-token cap - and only the model writing the rewrites was given room to think; that one configuration bounds how far their null reaches. That matters because the exploratory campaign's winning prompt (the first, on Section 3's 22-consultation pool) has as its central instruction to enumerate the transcript's facts and check them one by one, a procedure that takes room, and no candidate in any of the three searches ever had it: re-run unchanged on the held-out subset with a reasoning budget, it moves from 0.546 to 0.669 paired on omissions and acquires a usable flag rule (Section 7). So the nulls are consistent with a search confined to a budget in which the strategy it kept proposing could not run,

as much as with any limit of monolithic judging: a decision about how the evaluation was executed, invisible in everything the search reported, decided what it found. Two bounds stand: the wording null holds at the settings the searches scored under, and restoring the budget to an already-selected prompt is not the same experiment as searching at that budget, which we did not run.

Why: an omission leaves nothing to point at. An omission is AbsenceBench's problem at the evaluation layer: nothing for an evidence-seeking judge to cite. Every remedy above changes how hard the judge looks or how it reports, and none creates the thing it would look at; the optimisation campaigns make the same point from the other side - the search, free to write anything, wrote enumeration into its winning prompt.

7 WHAT WORKS: CONVERT THE ABSENCE QUESTION INTO PRESENCE CHECKS

One measurement among the reference judges points somewhere different. Standard evaluation tooling already contains a per-fact recipe - extract the source's facts once, check each against the text under evaluation, flag what is not covered. The RAGAS-style recipe, our own deliberately naive implementation of it in the lineage of RAGAS (Es et al., 2024), is unusable as shipped, checking roughly 62 unaudited facts per consultation and flagging 98.7% of clean notes. Its coverage component alone nonetheless reads 0.817 paired on omissions, past every monolithic judge on the same pairs. The omission signal appears as soon as the judge has a list of concrete facts; aggregating into a note-level flag destroys it.

Two routes complete the conversion and arrived at it separately (Figure C.1; Figure A.1 walks one case through both, and Table K.5 scores every rung). **The enumerate-then-check pipeline** builds it into the architecture, doing at judging time what construction did: extract the clinically material facts from the transcript alone, never from the note, so blindness is structural rather than promised, then verify each fact against the note as a closed lookup. **The two-stage pipeline** (B2 in the released data and tables) stops there, one keyed present-or-absent call per note; **the three-stage pipeline** (B3) adds a critic audit that drops unsupported or duplicated facts, grades each survivor's severity against the rubric of Section 3, and runs a quote-verified second look over anything flagged absent. Neither fixes a flag threshold in advance; every operating point below is swept and quoted with its own false-alarm rate. Fact enumeration for omission measurement is not new, MED-OMIT built it for medical summaries (Schumacher et al., 2023); what has not been shown is that it recovers judge detection where every monolithic remedy fails, and what to do with the verdicts. **The evolved prompt** needs no pipeline machinery: it is the exploratory campaign's optimiser-written winner from Section 6, whose central instruction is that same procedure, re-run unchanged at high reasoning effort under a 12,000-token cap, three times with different seeds.

7.1 HELD OUT, ENUMERATION WORKS, AND A BLINDED PHYSICIAN AGREES

On the 151 held-out pairs (131 omissions, 20 commissions, 47 clean twins over 47 consultations), the two-stage pipeline reads 0.786 paired on omissions, a mean of three runs spanning 0.771 to 0.801, against 0.531 to 0.649 for the best monolithic judge's three runs on the same pairs; replicate one's difference has a bootstrap 95% CI of [0.14, 0.27] (sign test $p = 0.002$). The paired figure replicated from the exploratory pass, 0.828 to 0.786; the single-note figure halved, from 31.2% at 9.4% false alarms to the 16.0% at 8.5% we report.

The strongest objection is circularity: the pipeline resembles the benchmark's own construction, so its win could be alignment between construction and judging rather than a property of the task; three things answer it. The RAGAS-style recipe, which shares no prompt text with the pipeline and extracts its facts with the judge model rather than the model that built the answer key, reaches 0.801 through its coverage component on the same pairs, the top of the pipeline's own three-run range, and concurrent work elsewhere reports the same monolithic null on a different clinical task (DeLucia et al., 2026). The extraction instrument, run blind, recovers 99.4% of facts authored by humans before any pipeline existed, on the sheets its own critic panel kept. And a physician author, shown ten notes on which the three-stage pipeline and the best monolithic judge reach opposite conclusions - blind to which judge said what and to whether the note had been edited, abstention offered - sided with the pipeline 10 of 10 ([72.2, 100], $p = 0.002$; six are the pipeline naming the removed fact, the

other four the three monolithic flags and one clean-note flag from its rule (Section 7.3)), a conditional result on selected disagreements with an author as the rater; the draw and design are Appendix E's.

7.2 What creates the signal: the list is about a third of it, the closed verdicts the rest

Two controls locate the mechanism, separating two ideas that sound alike. G-Eval (Liu et al., 2023) *decomposes the judgement*, splitting scoring into abstract criteria but still asking each as an open question over the whole note, and performs like the monolithic judges, 0.568 on omissions. The checklist judge goes halfway to *enumerating the facts*, and on the evaluation set falls between at 0.647, against the monolithic best of 0.634 and the RAGAS-style recipe's roughly 62-fact coverage component at 0.817. So the signal comes from listing concrete facts to check rather than from splitting the judgement into criteria, and a fuller list helps only up to a point: the recipe's roughly 62 facts and the two-stage pipeline's roughly 39 land on the same held-out paired score (Section 7.1), so beyond a reasonably complete list, length stops mattering.

The control that separates the two: hand the same monolithic judge the pipeline's own audited fact list, keeping the single open score. Paired omissions move from 0.570 to 0.634 in one held-out run (a held-out figure whose digits coincide with the evaluation-set best), against a control whose own three runs span 0.508 to 0.603 - a gain that stops inside the best monolithic judge's own held-out range (Section 7.1). Read through **a closed per-fact output structure** instead - a present-or-absent verdict keyed to each fact rather than one score over the note - essentially the same facts reach 0.786. Of the gap between a plain monolithic judge and the pipeline, having the list accounts for roughly a third and the closed verdicts for the rest (the fact block is itself a prompt change, so part of that gain could be wording). Even holding the list, the judge compresses all 198 held-out notes onto four distinct score values.

Deliberation is a real lever, and it is not the mechanism. Every control so far holds the judge's compute fixed, so the difference could be thinking time rather than task structure. Giving the monolithic prompts a reasoning budget and changing nothing else (Table B.2) lifts the plain completeness-scored judge from 0.570 to 0.670 paired on omissions, $+0.101$ $[+0.048, +0.153]$ and monotonic in dose, and gives **the engineered judge** - the strongest single-prompt judge we could write by hand, with eight worked examples - nothing at all, $+0.013$ $[-0.047, +0.071]$. A monolithic judge holding both advantages at once, the audited fact list and a high reasoning budget, reaches 0.699, still about 0.09 below the two-stage pipeline on the same pairs from a list of the same kind (the audited 37.6-fact list against the two-stage pipeline's 39.2 unaudited facts), at about three times the cost per note of its checking stage: compute closes roughly half the distance and then stops. Paired discrimination on added and altered content rises with the budget for every prompt given one, so the asymmetry widens with compute rather than closing (a direction rather than a size; Table B.2).

7.3 Act on the verdicts, not a score

The three-stage pipeline produces a graded verdict per fact and still needs a note-level decision, and averaging the verdicts into a coverage score is the weak way to make it: clean notes average 0.953 coverage and notes with omissions 0.929, so every threshold cuts inside a 2.4-point sliver, because one missing fact among about 38 barely moves an average. The usable object is **the per-fact rule**, a rule over the individual verdicts: *flag the note if any fact graded critical is judged absent*. Held out, it catches 20.6% of notes with an omission (27 of 131, [13.8, 28.2]; the lowest of three replicates, Figure K.3) and 32.8% of those carrying a critical-severity omission (19 of 58), at a false-alarm rate of 2.1% (1 of 47, [0.4, 11.1]) on the same single clean note in all three runs. The separation from its own noise is significant with the consultation as the unit of inference ($p = 2\text{e}{-06}$) and holds on the uniformly built omission subset alone (16.7%, $p = 0.008$). At that same noise level the best threshold on these *same verdicts'* aggregate score manages 7.6% (Figure K.3): the rule delivers 2.7 times the detection from identical evidence, and nothing was added to the judge, only the way its output is read. It was formulated post hoc from verdicts already recorded, then checked on the earlier exploratory pass, whose verdicts predate it, holding at 25.0% at 6.2%.

Over the full 495-pair evaluation set, whose prompts were never optimised on anything, though 47 of its 112 consultations are the held-out subset the rule was formulated on, the rule catches 24.6% of notes with an omission (72 of 293, [20.0, 29.8]) at a 2.7% false-alarm rate (3 of 112 clean twins,

[0.9, 7.6]) and 39.7% of the omissions graded critical (60 of 151); the 162 omission pairs it was not formulated on read 27.8% on their own. Below 5% false alarms exactly one aggregate threshold in the primary family clears its own noise, the engineered judge's 8.4% at the same 2.7% ($z = 2.88$), and the rule detects 2.9 times as much at an identical rate (Table K.5). It needs a severity grade on every fact *before any note is judged*, which the cheaper pipeline never produces, so the two are costed rather than ranked: the two-stage pipeline holds the higher paired score (0.786 against 0.762 as three-run means) at one fifth of the cost, while only the three-stage pipeline can express the rule. The rule's quietness rests on the audit stage's critical threshold, confirmed by the physician author on 1 of 10 rule-fired facts blind and 6 of 10 reconciled - the rater's lean, not the rubric's - so a clinician-aligned threshold would move the rate either way, on fewer facts or more (Appendix E).

7.4 The evolved prompt, the head-to-head, and real scribe output

Given room to execute its own instruction, the evolved prompt scores 45, 46 and 45 of the 47 held-out clean twins at exactly 10 out of 10 across three runs, turning "flag anything below 10" into a statable rule. On the full evaluation set - a fair test for this prompt too, its development pool carved out entirely - the majority of three catches 36.9% [31.5, 42.5] of notes with an omission at a 6.2% false-alarm rate and 58.9% of the critical-severity ones (89 of 151), against the per-fact rule's 24.6% at 2.7% and 39.7%. That threshold is a calibration, read off the first replicate's score distribution and held unchanged across the other two seeds and the full evaluation set, not a design guarantee: it works only because clean notes score exactly 10, which softens on the full set (hence 2.1% false alarms held out against 6.2% on the full set; Appendix J), so a deployment must re-establish it on its own clean notes. The calibration transfers across judge families: on the second family the same prompt scores 43 to 45 of the 47 held-out twins at 10 and reads 34.4% detection at 6.4% false alarms.

Collapsed to consultations as Section 4 requires, the evolved prompt detects significantly more notes with omissions at its own operating point ($p = 0.002$), firing on 25 consultations the per-fact rule misses against 7 the other way (held out the same contrast read 10 to 3, $p = 0.09$). Equalising the estimators (the evolved prompt is a majority of three runs, the rule one run) strengthens it. Neither point dominates and neither can be tuned to the other's: most of the evolved prompt's detection sits on notes scored exactly 9, so its next quieter threshold collapses to 14.3% at 0.9%, while widening the rule's trigger to supporting-severity facts lifts detection to 59.0% but fires on 35.7% of clean notes. The flags are different objects too, a bare score below 10 against the missing fact by name with its severity grade and the transcript quote behind it, so the pipeline buys precision and flag content rather than volume ($0.45 a note against $0.046, at one note per consultation).

On real scribe output, every operating point moves. Pointing the best of the eight and both routes, each at its own published rule, at 261 notes from the companion census (Fox et al., 2026) (all 87 carrying a panel-verified omission, and 174 the panel cleared) roughly doubles detection on both and raises false alarms 8.3-fold for the per-fact rule and 7.1-fold for the evolved prompt (Appendix J): a real note is less complete than a constructed pair everywhere rather than at one planted site. No threshold set on constructed clean twins survives the move, and the evolved prompt's flag-below-10 calibration fails as predicted, only 55.8% of panel-cleared notes scoring 10. Yet once thresholds are re-established there it still detects more than the monolithic judge at half its false-alarm rate (32.2% at 5.2% against the monolithic judge's 29.9% at 9.8% at its own rule), and the pipeline's flags keep their content: of the 45 notes it flagged among the 87, 34 (75.6% [61.3, 85.8]) name the very fact the census's panel verified. That match is a lexical-containment screen rather than an adjudicated precision, and the false-alarm reads are upper bounds, the cleared notes being clean only to the strictness of the census's own discovery (Appendix J).

7.5 The cost on commissions, and the open problem

The recovery has two costs. The first is **a cost on commissions**, and it belongs to the pipeline. On the evaluation set the enumerating designs sit at the bottom of the study on paired commissions - 0.744 down to the three-stage pipeline's coverage score at 0.661 (0.617 held out; Table K.5) - below every one of the eight designs while topping the omission ordering, and the inversion is at least as wide on the second judge family - mechanically what a coverage score should do, since a fabrication does not reduce coverage. One case bounds the mechanism: the engineered judge, a monolithic call with a prompt entirely about omissions, sits in the same band at 0.754, so part of the cost may be

attention. The evolved prompt does not pay it, holding paired commissions at 0.978 while detecting omissions, but its own cost is the per-deployment calibration; evaluation built on a coverage-scored pipeline needs a faithfulness judge alongside.

The second cost is shared: even the best method catches barely more than a third of notes containing an omission at a usable false-alarm rate, and the notes they miss are not a random sample. One case defeats every deployable method here. Restatement-trace omissions - the primary statement gone while an explicit or closely paraphrased restatement survives elsewhere - are 57 of the 293 omissions, the class in which the least information is lost, so pooling it into the headline numbers makes them conservative. Within the class every method in the primary judge family sits at or below its own noise floor (Figure K.4a) and the per-fact rule flags 0 of 57, while the second-family checker discriminates restatement traces well above chance (0.636 paired) and the rule still catches none: signal exists at the fact level and the decision layer cannot act on it.

The obvious clinical objection: a note that still says the thing somewhere is not deficient, so a quiet judge is right rather than blind. Splitting the same stored verdicts by where the survivor sits answers it. In 55 of the 61 restatement pairs corpus-wide (90.2% [80.2, 95.4]; 57 of the 61 sit in the evaluation set), the surviving restatement is under a different heading from the removed instance, and two of the six same-heading pairs are notes with no headings at all (Figure K.4b). The contrast with fragment traces is the mechanism: a fragment belongs to the same sentence or its neighbour and stays put (49 of the 87, 56.3%, under the same heading). What survives is therefore recoverable only by reading the note end to end: the fact is in the note, but not where the note claims to record it. Two qualifications belong beside that: in 37 of 61 pairs the survivor sits in the assessment-and-plan block, so a reader who works the whole note can typically still find the fact, and at these counts no detection gradient could have shown. The natural attack is structural rather than a better judge: score coverage per section instead of per note, turning the cross-section survivor from an undetectable absence into a detectable misplacement; the released site map supplies the labels. We state it as the field's open problem.

8 LIMITATIONS

Four bound the results directly; Appendix L states the full set. **The omission arm has no external anchor**: the only physician-labelled omission benchmark whose task matches ours is drawn from a corpus that is one of our own strata, so the omission results are internal to the instrument this paper releases. **One judge model family carries the headline results.** The second-family checks reproduce the asymmetry, the restatement-trace collapse and the pipeline's lead (Appendix G), and the evolved prompt's calibration transfers (Section 7.4), but single-note detection is judge-dependent, so those claims are scoped to the family measured; reasoning effort is not matched tier-for-tier across families (§5). **Errors are injected rather than naturally occurring**, and we measured the gap on both arms: paired 0.827 on MEDEC's physician-curated commissions against 0.939 on ours, and on omissions §7.4's real-note run, which moves every operating point; each pair also carries one planted error where a real note carries several (the pipeline judges 2.0 facts absent per real note against 1.1 on the benchmark's), so how failures interact is unmeasured. **One clinician validated this paper's own items, and endorsement is not blinding.** A physician author validated 94 items across two sittings - 70 across six stages of the benchmark and pipeline, and 24 more on the audit stage's own severity grades - with only three stages structurally blinded; they designed the instruments they graded, and every stage holds 6 to 20 items, so the intervals are the result, not the point estimates. The one independent check is of the severity rubric (Section 3), not of a judgement here.

9 CONCLUSION

Judges verify presence, not absence, and no remedy we tried creates usable detection in place: wording, criterion scope, voting and prompt search move the operating point along the same weak curve, and a larger reasoning budget lifts discrimination without closing the asymmetry. What works is converting absence into presence checks, with most of the recovery coming from the closed per-fact verdicts and the decision rule over them, not the fact list. Two methods implement it independently and trade off rather than rank, and an omission whose restatement survives elsewhere defeats both; we release the graded benchmark as the instrument for attacking it.

ETHICS STATEMENT

The corpus contains no real patient data. PriMock57's consultations are between clinicians and actor patients (CC BY 4.0), ACI-Bench's encounters are simulated doctor-patient conversations released for research (CC BY 4.0), and the remaining strata are authored scenarios; no patients were involved and no clinical records were used. The three commercial scribe products whose notes appear in the companion census (Fox et al., 2026) and in Section 7.4's real-error run are anonymised throughout as Scribe A, B and C, and the text of the notes they wrote is not released: the products' terms of service differ on republication, and we withhold all three alike rather than release asymmetrically, which would make products identifiable by their absence (Appendix F); released findings carry the transcript-side evidence and quote the note's own words only in the ten internal-contradiction findings that cannot be stated without them. Nothing in this paper is a clinical-safety evaluation of any named product, and no result here should be used to select or reject a scribe for clinical use. The authors build evaluation tooling commercially, which is a competing interest: the pipeline this paper recommends is the authors' own design, and the reference set includes a judge from the authors' production evaluation; every judge evaluated here is fully specified in the release, prompts, model pins and judgements included, so any reader can check the comparison.

REPRODUCIBILITY STATEMENT

A public release accompanies this paper, withheld from the submission only because a live link would identify the authors and could expose reviewer traffic; the links are published when the submission is de-anonymised. It holds the 500 graded pairs with their clean notes, errored twins, removed facts and grades; the fact-site map and both graders' severity verdicts; the transcripts and fact sheets; every verified census finding with products anonymised; every benchmark judge run's judgements with the raw completions (the real-vendor-note run's records judge withheld note text and are withheld with it); the clinician-validation records, as keys, answers and scored results (the item packs that print vendor note text are withheld with it; the two that carry none - the severity items with the rubric verbatim, and the construction-adjudication queue - ship); and the full harness. Appendix F is the release index and Appendix D pins every model, setting and seed. The construction protocol is Appendix A, the pipeline's deployment parameters Appendix C, the three optimisation campaigns with their seed and winning prompts verbatim Appendix B, and the second judge family in full Appendix G.

USE OF AI

Generative AI tools were used to aid editing parts of this paper, and to write and edit the software behind it. The benchmark's note pairs and fact sheets are produced by LLM steps as part of the method itself, each verified as Section 3 and Appendix A describe, and the severity grading and audit critics are likewise LLM instruments; we note them here because ICLR's disclosure list names synthetic data generation, though they are the study's method rather than assistance with it. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026. Oral presentation. arXiv:2507.19457.

Taylor N Anderson, Vishnu Mohan, David A Dorr, Raj M Ratwani, Joshua M Biro, and Jeffrey A Gold. Evaluating the quality and safety of ambient digital scribe platforms using simulated ambulatory encounters. *Mayo Clinic Proceedings: Digital Health*, 3(4):100292, 2025. doi: 10.1016/j.mcpdig.2025.100292.

Suhana Bedi, Bridget Lin, Anson Y. Zhou, Chloe O. Stanwyck, Jenelle A. Jindal, Sanmi Koyejo, David Stutz, and Nigam H. Shah. CARE: A conformal safety layer for medical summarization. *arXiv preprint arXiv:2606.08969*, 2026.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. MEDEC: A benchmark for medical error detection and correction in clinical notes. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22539–22550, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1159. URL https://aclanthology.org/2025.findings-acl.1159/.

Henry Bergman, Vivian Liu, Ben Austin, and Rohan Sangera. Are automated documentation-error judges fit to measure ambient AI scribes? A pre-registered, blinded human-validation study. medRxiv preprint, 2026a. Posted 17 August 2026.

Henry Isaac Bergman, Vivian Liu, Ben Austin, Sarah Ali, Merle Fiedler, Celene Sandiford, Rebecca Blanchard, Carrlos Lopez Casanovas, Giulia Pedrazzini, Theo Markopouliotis, and Ferdinand Vermersch. Quality, consistency, and clinical safety of AI-generated versus clinician-written clinical notes: a multi-country paired simulation study. medRxiv, 2026b. Preprint, posted 21 August 2026; all authors affiliated with the vendor evaluated.

Joshua Biro, Jessica L Handley, Nathan K Cobb, Varsha Kottamasu, Jeffrey Collins, Seth Krevat, and Raj M Ratwani. Accuracy and safety of AI-enabled scribe technology: Instrument validation study. *Journal of Medical Internet Research*, 27:e64993, 2025a. doi: 10.2196/64993.

Joshua M Biro, Jessica L Handley, J Malcolm McCurry, Adam Visconti, Jeffrey Weinfeld, J Gregory Trafton, and Raj M Ratwani. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *npj Digital Medicine*, 8:222, 2025b. doi: 10.1038/s41746-025-01586-2.

Wenhui Chen, Jianlin Chen, Ziyao Lin, Peiji Long, and Chi Man Vong. Judging is not enumerating: Silent omissions in LLM-authored acceptable sets, 2026. URL https://arxiv.org/abs/2608.01000.

Alexandra Dahlberg, Tiila Käenniemi, Tiia Winther-Jensen, Olli Tapiola, Rami Luisto, Tuukka Puranen, Max Gordon, Enni Sanmark, and Ville Vartiainen. Measuring the quality of AI-generated clinical notes: A systematic review and experimental benchmark of evaluation methods. *Artificial Intelligence in Medicine*, 177:103421, July 2026. doi: 10.1016/j.artmed.2026.103421.

Alexandra DeLucia, Heyuan Huang, Sonal Joshi, Mahsa Yarmohammadi, Ahmed Hassoon, and Mark Dredze. Same verdict, different reasons: LLM-as-a-judge and clinician disagreement on medical chatbot completeness. *arXiv preprint arXiv:2604.16383*, 2026.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In Nikolaos Aletras and Orphee De Clercq (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, St. Julians, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Sebastian Fox, Luke Markham, Ryan Lail, and Michael Karotsieris. One note in three: a verified census of three deployed AI scribes, and the instrument that counted it, 2026. URL https://arxiv.org/abs/2608.31017.

Harvey Yiyun Fu, Aryan Shrivastava, Jared Moore, Peter West, Chenhao Tan, and Ari Holtzman. Absence Bench: Language models can't see what's missing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight poster. arXiv:2506.11440.

Khizar Hussain, Bradley A Malin, Zhijun Yin, Susannah Leigh Rose, and Murat Kantarcioglu. Blending human and LLM expertise to detect hallucinations and omissions in mental health chatbot responses. *arXiv preprint arXiv:2604.06216*, 2026.

Annessa Kernberg, Jeffrey A Gold, and Vishnu Mohan. Using ChatGPT-4 to create structured medical notes from audio recordings of physician-patient encounters: Comparative study. *Journal of Medical Internet Research*, 26:e54419, 2024. doi: 10.2196/54419.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. PriMock57: A dataset of primary care mock consultations. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 588–598, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.65. URL https://aclanthology.org/2022.acl-short.65/.

Elliot Schumacher, Daniel Rosenthal, Varun Nair, Luladay Price, Geoffrey Tso, and Anitha Kannan. MED-OMIT: Extrinsically-focused evaluation of omissions in medical summarization. *arXiv preprint arXiv:2311.08303*, 2023.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6594–6604, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.529. URL https://aclanthology.org/2021.emnlp-main.529/.

Hui Shen, Tianyang Wu, Fei Wang, Jin Fang, Yuange Li, Xiaoling Wu, Shuyi Liu, Liting Chen, Qiuping Ren, Xiao Meng, Jiatong Xu, Jie Sun, Yujie Zhao, Xin Liu, Liaoyuan Wang, Guipeng Mai, Jingjing You, Zhe Jin, Xuewei Wu, Wenle He, Xue Han, Shuixing Zhang, Dong Zeng, and Bin Zhang. Error detection in emergency radiology reports using a large language model: Multistage evaluation study. *Journal of Medical Internet Research*, 28:e86841, 2026. doi: 10.2196/86841.

Sandra L Taylor, Melissa Jost, Scott MacDonald, Yunyi Ren, Shelley Hilton, Sadie Davenport, Debbie Aizenberg, Bruce Hall, Courtney R Lyles, and Jason Y Adams. Quality of clinical notes created by ambient listening generative AI: Pragmatic prospective pilot study. *JMIR Medical Informatics*, 14:e86474, 2026. doi: 10.2196/86474.

Bhavik Vachhani, Kush Shrisvastava, Pranshu Nema, and Sai Chiranthan. Beyond literal summarization: Redefining hallucination for medical SOAP note evaluation, 2026. URL https://arxiv.org/abs/2604.14829. Vendor-authored preprint (institutional byline: a clinical documentation vendor's research group).

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, 2023. doi: 10.1038/s41597-023-02487-3.

Karen Zhou, John Michael Giorgi, Pranav Mani, Peng Xu, Davis Liang, and Chenhao Tan. From feedback to checklists: Grounded evaluation of AI-generated clinical notes. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1485–1499, Suzhou (China), November 2025. Association for Computational Linguistics. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.104. URL https://aclanthology.org/2025.emnlp-industry.104/.

A GROUND TRUTH: CONSTRUCTION PROTOCOL, INSTRUMENT TRAJECTORY, AND THE SEVERITY RUBRIC

This appendix is the construction record behind Section 3. It gives the fact-sheet schema and the critic panel that audits it, the blind-recovery check and the sheet-size asymmetry it exposed, the severity rubric printed in full, the reference-note repair loop, the pair-construction rules and their verification outcomes, and the two cohorts of pairs carried over from an earlier build. A replicator should be able to rebuild the benchmark from this appendix, the released prompts and the released artefacts. Model pins and seeds for every stage are in Appendix D, with the transport and hash records in the released harness; Figure A.2 is the same pipeline as a schematic.

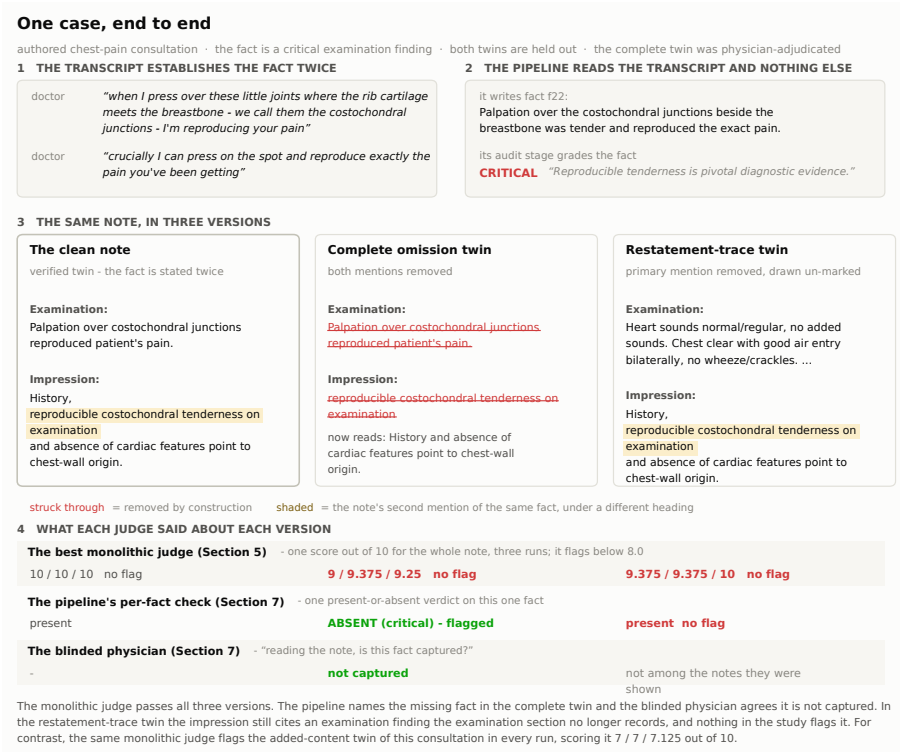


Figure A.1: One case end to end: a critical examination finding the transcript establishes twice, the note records twice, and two twins remove differently. The monolithic judge passes all three versions; the pipeline's per-fact check names the missing fact in the complete twin and the blinded physician agrees it is not captured. The restatement-trace twin is drawn un-marked, as a reader sees it: the examination reads normally, the impression still cites the finding it no longer records, and nothing in the study flags it. The case is an authored consultation, so the figure excerpts only our own transcript and notes and no vendor text appears anywhere.

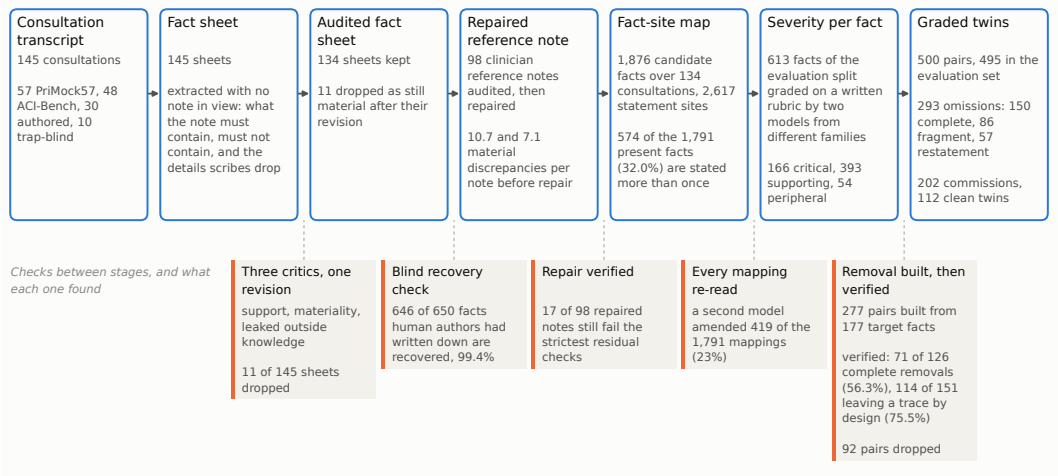


Figure A.2: The answer key is built from the transcript, not borrowed from the reference note. Each box is a construction stage and carries what it produced; each chip below is the check that stands between two stages, and what that check found or rejected. Severity is graded per fact by two models from different families on a written rubric; they split on 93 of the 613 facts, and the lower grade is taken. The two checks that carry findings of their own are blind recovery, where extraction recovers 646 of 650 facts that human authors had written down for the same consultations, and removal verification, where only 71 of 126 attempts to remove every trace of a fact from a real note survive a cross-family panel against 114 of 151 for removals that leave a trace by design.

A.1 THE FACT SHEET AND THE THREE CRITICS

A.1.1 THE SCHEMA

For each consultation we extract a fact sheet from the transcript alone, in one call, with no note of any kind in the prompt. The sheet has exactly three keys.

key	one item is	fields
`must_contain`	one atomic fact a correct and complete note of this consultation has to include	`fact` (precisely stated, no compounds), `evidence` (a verbatim transcript quote of at most 25 words that supports it), `load_bearing` (`high` or `medium`, how much a competent clinician's reading of the note depends on it)
`must_not_contain`	one specific wrong claim a scribe could plausibly make on this consultation	`assertion` (the erroneous claim), `why_wrong` (one line on what the transcript actually establishes)
`salience_traps`	one moment in this transcript where a scribe could plausibly go wrong	`trap` (what is tempting to get wrong), `correct_handling` (what a correct note does), `mode` (one of omission, negation, dose value, laterality, attribution, modality hardening, temporal, fabrication, decision status, anchoring), `importance` (the severity grade of A.3)

The extraction prompt directs coverage at the clinically decisive content of a consultation: presenting complaint, key positives and pertinent negatives, examination findings or the explicit absence of an examination on a remote consultation, the working impression with any hedging preserved, the plan with drug names, doses, frequencies and durations as actually said, safety-netting, and material

social or risk history. Three rules do the work that makes the sheet usable as an answer key: one fact per item with no compounds; nothing enters `must_contain` without a quotable supporting span; and where the transcript is ambiguous the sheet preserves the ambiguity rather than resolving it. The `evidence` and `load_bearing` fields exist so that a support audit can be run mechanically rather than by re-reading the transcript for every item. Both prompt versions are released in the code repository as `prompts/extraction_prompt_v1.txt` and `prompts/extraction_prompt_v2.txt`.

A.1.2 THE THREE CRITICS

Every extracted sheet is audited by three separate critic calls before it is used for anything. The transcript is immutable, so the panel asks only whether the sheet is a faithful reading of it.

- **Support** checks internal consistency between transcript and sheet: every `must_contain` item genuinely said, every trap a real moment in the dialogue, every `must_not_contain` item really an error for this consultation, no self-contradiction.
- **Materiality** judges clinical importance rather than support: it flags trivia in the required facts (padding raises the completeness bar and penalises good notes), clinically important transcript content missing from `must_contain` (a working diagnosis, a red-flag safety net, a drug dose, a stated allergy, a decisive negative), and `load_bearing: high` grades clearly wrong in either direction.
- **Leakage** guards the one contamination channel a blind extraction has: any item resting on information not in the transcript - outside clinical knowledge asserted as fact, content that could only come from a reference note or from knowing how scribes typically fail, meta language, and paraphrase drift where the `evidence` quote does not appear in the transcript or does not support the `fact`.

Critic findings carry a severity of `minor` or `material`. Any material issue triggers **one revision cycle**: the sheet is rewritten to fix only the flagged issues, with the transcript required back character-for-character identical, and the revised sheet is re-audited by the support critic. There is no second revision. A sheet whose material issues survive the revision is dropped, not patched, with the one exception described in the released instrument-trajectory record.

A.1.3 WHAT THE PANEL KEPT

stratum	consultations	sheets kept	dropped
PriMock57 (real UK primary-care recordings, clinicians with actor patients)	57	53	4 (7.0%)
ACI-Bench (US ambulatory, subsampled)	48	45	3 (6%)
Trap-blind authored (a stratum written with no knowledge of the trap schema, as a representativeness guard)	10	9	1 (10%)
Authored consultations, re-extracted blind	30	27	3 (10%)

That is **134 audited sheets** (53 + 45 + 9 + 27 blind re-extractions), alongside the 30 authored consultations' own sheets, written when those consultations were written, which are what A.2's blind-recovery check scores against. Counted as consultations rather than as sheets, 137 distinct ones carry a sheet. The 134 consultations that carry a verified clean note (A.5, F.2) are a different set of the same size as these 134 sheets. The ACI-Bench sample of 48 is drawn from 207 encounters with 67 excluded because they had been used to train a prior optimised judge in this codebase; those 67 turn out to be exactly ACI-Bench's official training split, so the sample is drawn purely from data held out from that judge (`provenance/subsample_manifest.json` in the code repository).

A.2 BLIND RECOVERY OF AUTHORED GROUND TRUTH, AND THE SHEET-SIZE ASYMMETRY

The instrument that writes the answer key needs its own accuracy measured, and the 30 consultations we wrote ourselves make that possible: each already carries the fact sheet its authors produced alongside it, so re-extracting a sheet blind from the transcript alone and comparing it against that authored sheet is a direct test.

Recovery is 646 of 650 authored required facts, 99.4%, cluster bootstrap 95% interval [98.8%, 99.9%], with a **per-consultation minimum of 93.8%** - so the recovery is not an average hiding one bad consultation. The 650 facts are those of the 27 consultations whose blind re-extractions passed the critic panel (of the 30 re-extracted; the three drops are in A.1's table). The reading is that blind extraction is a faithful substitute for an authored sheet, that the extracted strata are not systematically thinner than the authored one, and that pooling the strata in the main analysis is licensed.

The size comparison went the other way, and we did not anticipate it.

per consultation	authored sheet	blind-extracted sheet
required facts	24.1	**42.4**
forbidden assertions	16.1	**24.0**
salience traps	7.0	**17.5**

Extracted sheets are 1.8 times fatter on required facts and 2.5 times fatter on traps. That is a size difference between strata, and it cuts against our own hypothesis rather than for it: a higher completeness bar on the extracted strata means a judge could appear to miss more omissions there for reasons that have nothing to do with whether judges can see absence. We report it as it stands.

The consolidation pass narrows it. Every extracted sheet's required facts were classified as `core` (a competent clinician would consider the note deficient without it) or `contextual`, with the classifier never shown a target count and no curving of any kind. Pairs and clean notes are built from the core view only; the full sheet is retained alongside it and nothing is deleted. Per-stratum mean core counts against the authored sheets' 24.1: **PriMock 24.8**, ACI-Bench 29.0, trap-blind 31.1, authored re-extraction 32.6. The PriMock stratum lands on the authored envelope; the others narrow substantially but stay fatter, and that is reported as it fell. On traps, **2,154 of 2,296 traps are eligible for use as pair targets**, with 142 peripheral-importance traps excluded from injection.

The reverse check found gaps in the human sheets: asking which core facts the blind extraction found that the authored sheets had missed gives 13 gaps across 7 of the 30 authored consultations, mostly safety-netting and pertinent negatives. Blind extraction caught note-worthy content the human authors missed. This is reported as a result; the authored sheets and the 90 pairs built from them are left untouched rather than repaired, so nothing downstream shifts because of it.

A.3 THE SEVERITY RUBRIC, PRINTED IN FULL

A.3.1 SEVERITY WITHOUT A RUBRIC

Grading an omission by how much it matters is only useful if the grade is stable. Natively it is not. Two frontier models - successive generations of one model family - grading the same 171 semantically matched traps independently agree at **Cohen's kappa 0.177**, cluster bootstrap 95% interval [0.04, 0.31], where 0 is the agreement expected by chance alone. Raw agreement is 66%, which sounds tolerable until the chance baseline is taken out. The disagreement is one-sided rather than noisy: of 52 traps that one generation graded `supporting`, the other graded **37 as `critical`**. The kappa implementation was checked against perfect-agreement, systematic-disagreement, textbook and independence cases before it was used on real data.

A.3.2 SEVERITY WITH THE RUBRIC

We wrote a clinical rubric with three grade definitions, worked anchors and a four-step decision procedure (printed below; it also ships verbatim, with its hash, in the released `validation/sitting_severity_items.json`), and re-graded **all 683 traps (211 authored plus 472 extracted)** with two independent model graders from different families.

- **Cross-family agreement with the rubric - the two graders whose consensus the benchmark ships - is kappa 0.662**, cluster bootstrap 95% interval [0.59, 0.73], against **0.177 for the unanchored same-family pair**, with non-overlapping intervals.
- The gain is not an artefact of the easier design: the original, harder cross-sheet matched-pair design still reads 0.50 to 0.56 under the rubric, with about 0.16 of the residual gap shown by control to be item-matching noise.
- **The rubric moves one family's own grades substantially**: kappa 0.39 between that family's rubric-anchored grades and the earlier generation's unanchored run. The rubric is doing real work on how the model grades, not relabelling grades it would have produced anyway.
- **Both rubric-anchored graders grade more severely than either model does natively.** On the 472 extracted-sheet traps graded by both runs, one family's rubric run gives 302 critical, 158 supporting and 12 peripheral, against 243, 190 and 39 for the same family's native grading - 64% critical against 51%. The action and safety tests fire often. We disclose this wherever a result is conditioned on severity: the paired finding is that severity is contested between frontier models until a written rubric anchors it, and that the rubric shifts the distribution towards critical.
- **92 of 683 remaining disagreements (13.5%) are resolved conservatively to the lower grade** and flagged in the data, because the second family is the more severe grader of the two. The consensus labels are the working severity axis.
- **The rubric has since been graded blind by an independent clinician, and that is the only check from outside the author group anywhere in this paper.** Independent here means not an author of this study and with no involvement in it at any stage; it is not a claim about how the clinician was found, and this paper does not describe that. The companion census (Fox et al., 2026) applies the same written rubric to its own verified findings, and a fresh sample of them was put to that clinician: **9 of 12 grades exact against the rubric and 3 one grade apart**. Read with the same census's earlier sitting by a physician author, **25 of 32 grades are exact and no disagreement anywhere exceeds a single grade**, so the rubric reproduces to within a grade under two clinicians who graded disjoint material. What the pair does not establish is a direction: the two clinicians lean opposite ways against the same rubric - three of twelve below it for the non-author clinician, four of twenty above it for the author - so no claim in this paper rests on the rubric leaning one way, and the one-grade gap Section 7.3 reports for the pipeline's audit-stage grades is read as a property of that rater as much as of the rubric. The check is of the rubric only. No adjudication of a judge output in this paper has been read by anyone outside the author group, and Section 8 states that gap.

A.3.3 SEVERITY AS IT ENTERS THE BENCHMARK

The same rubric, run verbatim, grades the facts the pairs are built from. The evaluation split carries **613 graded facts: 166 critical, 393 supporting, 54 peripheral**, from two cross-family graders on the verbatim rubric, with disagreement taking the lower grade and flagging the fact. **The two graders disagreed on 93 of 613 facts (15.2%)**. A later pass graded the optimiser's development split as well, so the full severity artefact holds **880 facts (243 critical, 558 supporting, 79 peripheral, 137 flagged = 15.6%)** at **$44.19 over 1,760 calls**.

Peripheral facts are scarce in this corpus. Under a rubric whose third test is whether losing the fact materially weakens the note as a clinical record, almost everything a scribe records earns at least `supporting`. **54 peripheral facts out of 613**, and **36 of those 54 come from contextual required-fact items** - exactly the class the core view drops - which is why every earlier build of the benchmark contained no peripheral pairs at all.

Where a trap already carried an importance grade from its own fact sheet and was also graded fresh under the rubric, the two agree on 113 of 158 trap facts (71.5%), with disagreements running both ways; the per-fact record is in the released severity artefact (`pairs/factorial_severity.json` in the data repository).

A.3.4 THE RUBRIC, VERBATIM

The following is the rubric as the graders received it. The file's administrative header (internal file references and an author name) is omitted, and one internal workstream code in the final sentence is replaced in brackets; the graded text is otherwise unaltered, and the released file carries the same redactions.

```
# Severity rubric - importance grading for salience traps and findings

## The question being graded

For a trap or finding, grade the clinical importance of the note
    getting this
RIGHT vs getting it wrong (omitting it, fabricating it, or altering it):
**if the note carried this error, how much would it matter for the
    patient or
the next clinician who reads the note?**

Grade the consequence of the error IN THE NOTE, not the drama of the
    topic.
A mundane-sounding fact can be critical (a drug allergy); a
    dramatic-sounding
one can be peripheral (a vividly described but self-resolved symptom
    with no
bearing on the plan).

## The three grades

**critical** - the error would plausibly change clinical action, delay
    or
misdirect care, or create a safety risk. Anchors:
- omitted working diagnosis or a hedged impression hardened into
    certainty
- omitted red-flag safety-netting ("return if the headache becomes
    sudden")
- omitted or wrong drug allergy, interaction, or contraindication
- wrong drug, dose, frequency, or duration; a conditional plan made
    definite
- fabricated examination on a remote consult; invented findings or
    results
- omitted pertinent negative that licenses the management decision (the
    denied
  cardiac features behind treating chest pain as musculoskeletal)
- wrong patient identity attributes (name, sex, age used clinically)

**supporting** - degrades the note's completeness, clarity, or
    defensibility,
but is unlikely to change what happens next. Anchors:
- omitted duration/onset detail that colours but does not gate the plan
- omitted social context relevant to follow-up (lives alone,
    occupation) where
  no immediate action hangs on it
- compressed phrasing that loses nuance without inverting meaning
- omitted secondary symptom that neither supports nor threatens the
    diagnosis

**peripheral** - no plausible clinical consequence. Anchors:
- conversational colour, rapport, patient-education phrasing
- administrative closure ("no further questions")
- redundant restatement of something already captured elsewhere in the
    note

## Decision procedure (apply in order)
```

```
1. **Action test**: would a competent GP, or the next clinician reading
    the
   note, plausibly DO something different if this error stood? Yes ->
    critical.
2. **Safety test**: even if action today is unchanged, does the error
    remove a
   safety net or create a latent risk (allergy, red flag, follow-up
    trigger)?
   Yes -> critical.
3. **Record-quality test**: does it materially weaken the note as a
    clinical
   record (completeness, defensibility, handover value)? Yes ->
    supporting.
4. Otherwise -> peripheral.

Tie-breaks: if genuinely torn between two grades after the procedure,
    take the
LOWER grade (conservative - protects the severity analysis from grade
inflation). Grade each item independently; do not curve to a
    distribution.

## What this rubric is for (and not)

It anchors model graders and the author-clinician to one written
    standard so
grades are comparable across models and time. It does not claim to
    settle
clinical importance in general - cross-grader disagreement UNDER this
    rubric
is itself a reported result [of this study].
```

The physician author's blinded validation of these grades is in Appendix E.

A.4 THE REFERENCE-NOTE REPAIR LOOP

The clean note that every pair is built from is a repaired version of the corpus's own clinician reference note. The repair runs per consultation, after that consultation's fact sheet has passed the critic panel, and it sees three things: the transcript, the audited fact sheet, and the original clinician note.

It works in two steps. **The audit** lists every discrepancy between note and transcript, classified as `missing_fact`, `contains_error`, `unsupported` or `hardened_uncertainty` (the transcript hedged, the note is definite), each graded `material` - could change patient understanding or management - or `minor`. **The repair** fixes every material discrepancy with minimal edits, preserving the note's structure and style so the result reads as the same clinician's note corrected rather than a rewrite; both versions stay in the released record. A verification pass re-checks the repaired note, and notes that fail are fixed once more and re-checked.

Every reference note in both public strata carries at least one material discrepancy with its own transcript: PriMock **53 of 53 (100%)**, Wilson 95% interval [93.2%, 100%], and ACI-Bench **45 of 45 (100%)**, [92.1%, 100%]. The density is what matters more than the bare 100%.

	PriMock (53 notes)	ACI-Bench (45 notes)
discrepancies, all severities	1,030	703
of which material	566	320
mean per note, all severities	19.43	15.62
mean material per note	**10.68**	**7.11**

By kind, with material counts in brackets: PriMock missing fact 574 (310), contains error 195 (139), unsupported 154 (60), hardened uncertainty 107 (57); ACI-Bench missing fact 356 (185), unsupported 193 (42), contains error 98 (57), hardened uncertainty 56 (36). Missing facts dominate on both. Graded on the rubric of A.3, PriMock's 1,030 discrepancies are 345 critical, 500 supporting

and 185 peripheral, and ACI-Bench's 703 are 143 critical, 374 supporting and 186 peripheral. Every discrepancy ships with its verbatim evidence quote in the released audit record (`provenance/wd_r3_report.json` in the code repository); Figure A.1 walks one case end to end.

Convergence. No note needed zero edits: the mean edit count is 18.74 per PriMock note and 15.69 per ACI-Bench note. Of the 53 PriMock notes, 19 passed verification first time, 34 needed a further fix and 23 of those then passed, leaving **11 still failing** the strictest residual checks. Of the 45 ACI-Bench notes, 18 passed first time, 27 needed a further fix and 21 then passed, leaving **6 still failing**. So **17 of the 98 repaired notes still fail verification on ultra-strict residuals**; all 17 are logged by identifier and none is silently kept.

Three framing points belong with these numbers. The 100% is by our audit instrument's standard, a single-model auditor applying a strict rubric, so we report the discrepancy density with its verbatim evidence quotes and its severity split rather than resting on the saturated proportion. PriMock's reference notes are terse case summaries by design, which inflates the missing-fact counts specifically. And the instrument doing the auditing is the same extraction machinery whose blind recovery of human-authored facts is 99.4% in A.2, so the checker has itself been checked.

A.5 Pair construction, the single-edit invariant, and what verification rejected

A.5.1 Four classes and what a single edit means for each

The corpus carries four pair classes. The older three-value `type` field (`add` / `change` / `omit`) is retained unchanged for compatibility, and `class` is the field that distinguishes the two kinds of omission.

class	the edit
`add`	one span asserting something the clean note does not support
`change`	one span's assertion altered
`omit-complete`	one fact removed from the note entirely, at every place it is stated
`omit-partial`	one fact's primary statement removed, its other statements left in place

The invariant that keeps a pair interpretable is that exactly one thing changed. What "one thing" means depends on the class, and getting that right was the hardest definitional problem in the build. For `add` and `change` the unit is a **span**: one non-equal block in the diff, with a blast-radius cap. For `omit-complete` the unit is a **fact**: removing a fact that lives in three sections is one edit by any standard a clinician would recognise and three edits by the diff, so the one-block rule is replaced by two conditions - every non-equal block is a deletion, or a repaired wound no longer than what it replaced, and every deleted span is a mapped statement of the target fact. For `omit-partial` the unit is a span again, the primary site, algorithmically identical to `add` and `change` and semantically its opposite, because here a residual is required rather than forbidden.

The **primary site** is the note's home statement of the fact: the site carrying it most completely and specifically, the one a reader would cite. It is nominated by the cross-family verifier, with recorded fallbacks to the mapper's nomination and then a deterministic order rule; which rule fired is recorded per fact. Every `omit-partial` pair carries residual metadata - the number of surviving mentions, each with its span, section and strength, plus the strongest surviving strength and the sections involved. Strength has three grades, strongest first: `explicit` (states the fact plainly in the note's own clinical terms), `paraphrase` (states the same fact in different words) and `partial` (states only part of it, losing a defining detail such as a value, a laterality, a hedge, or the half that drives action). There is deliberately no "implied" grade, because a span a reader could only infer the fact from is not a statement of it, and treating one as a site would delete unrelated clinical content.

A.5.2 THE ONE PERMITTED REVISION OF THE EDIT CONSTANTS

The algorithmic single-edit constants took their one permitted revision, exercised before any semantic edit check ran, and the revision is disclosed here in full because it changed which pairs the build could produce. The first version of the constants rejected **210 of 411 pairs as built (51%)**, including **20 of the 90 preserved authored pairs, 13 of them omissions** - the study's headline error class. The dominant false-failure classes were one-block contiguous section deletes of three to five units and adds or omissions realised as natural one-to-one line replacements, both of which are single-location edits under the rule's own stated intent. The second version raised the pure insert or delete cap from 2 to 6 units and the replace cap from 2 to 3 units a side, and dropped two proxies, word-operation purity and a maximum-regions count, because edit direction and atomicity are now verified semantically instead. Three things stayed frozen: the single non-equal block invariant that localises the edit, the 30-changed-word blast-radius cap, and the regeneration suffix's two-unit limit. The second version's fail count as built is **97**. No further revision is permitted, and none was taken.

A.5.3 THE SITE MAP

Neither omission class can be built without knowing where in the note a fact is stated, so a mapping stage runs before any injection. Candidate facts per consultation are the pair-eligible omission-mode traps first, then core required facts with `load_bearing: high`, capped per consultation and seeded on the study seed. For each candidate the mapper locates every place in the clean note that states it - the smallest verbatim span, its section, its strength, its character offset - and the map is built by one model family and checked by another, asking whether each listed span really states the fact, whether any site was missed, and whether the primary nomination is right.

Over **134 consultations (112 evaluation plus 22 in the optimiser's development pool)** at 14 candidate facts each: **1,876 facts mapped, 1,791 present in the note, 85 absent, 2,617 sites located**, at **$155.63 over 537 calls**.

574 of the 1,791 present facts (32.0%) are stated in more than one place, at a mean of 1.461 sites per present fact. Redundancy concentrates in the facts that matter clinically:

candidate bucket	facts stated more than once
supporting traps	71 of 116 = 61.2%
critical traps	80 of 155 = 51.6%
peripheral traps	8 of 23 = 34.8%
core required facts, high	276 of 813 = 33.9%
core required facts, medium	31 of 135 = 23.0%
contextual required facts	108 of 549 = 19.7%

The second call is not optional at this scale: the verifier **amended 419 of 1,791 facts (23%)** - dropping a span that did not state the fact, adding one the mapper missed, or moving the primary nomination - and 14 verification failures were excluded from construction. Of the 2,617 sites, 1,224 align to whole units and can be deleted with no model call at all. The full verdict counts, the per-stratum redundancy split and the sites-per-fact distribution are in the released site map (`pairs/fact_sites.json`).

A.5.4 ALLOCATION AND BUILD

The allocator ran to a target of 42 pairs for each of the nine trace-by-severity combinations over the 112 evaluation consultations, capped at 4 target facts and 6 pairs per consultation, giving **277 pairs planned from 177 target facts** (ACI-Bench 125, PriMock 83, authored 50, trap-blind 19). The build cost **$7.29 over 173 calls**, or **$0.026 a pair**, and the edit ledger totals: **104 pairs were pure programmatic span deletions with no model call at all, 65 needed a grammar tidy, 77 needed a collateral cleanup** where deleting a span orphaned a neighbouring clause (340 words dropped in total, every edit logged), and on **31 the attempted cleanup was rejected by the validator** so the raw deletion stood. Those four numbers sum to the 277 planned pairs. A grammar tidy is punctuation-level repair and nothing more: in one logged case, deleting the clause recording that vomiting came first left "onset Sunday evening.; nausea settling.", and the tidy made it "onset Sunday evening;

nausea settling." - no word of remaining content changes. Construction left **exactly one residual site on all 151 partial pairs**, by design.

A.5.5 VERIFICATION STANDARDS, PER CLASS

`add` and `change` pairs go through the unchanged edit check. The omission classes have their own standards.

class	semantic checks
`omit-complete`	`truly_absent` (the removed fact was present in the clean note and survives nowhere in the errored note, with no partial or paraphrased residual mention - verbatim the same clause the earlier build used, so the two are directly comparable); `spans_all_state_fact` (every removed span states the target fact, the other half of "one edit means one fact"); `no_collateral_loss` (every other fact the clean note recorded is still recorded); `reads_naturally` (the result reads as a note that never recorded the fact, not as a note visibly cut about)
`omit-partial`	`primary_site_removed` (the home statement is gone); `residual_present` (at least one mention survives, quoted and graded - the inverse of `truly_absent`, and where the residual metadata comes from); `no_collateral_loss`; `reads_naturally`

Free algorithmic checks gate all of it first (deletion-only diff, no new content, removed and residual spans as required, every mapped site of every other fact still present). One instrument detail matters for anyone reproducing this: the omission checks are given the errored note and **no transcript**, because every question is note-internal, and supplying the transcript made the auditor fail `truly_absent` on the grounds that the fact still appears in the transcript, which it always does.

Verification is a cross-family panel of two models, with a third call from the auditor at a different seed on any field the two split on and a per-field majority deciding: **679 calls, $33.92**.

class	verified	passed the algorithmic checks	panel unanimous	verified and unanimous
`omit-complete`	**71 of 126 (56.3%)**	121 of 126	90 of 126	60
`omit-partial`	**114 of 151 (75.5%)**	145 of 151	130 of 151	101

A.5.6 THE COMPLETE CLASS'S FAILURE BREAKDOWN

55 of 126 complete-removal attempts (44%) failed verification, and the breakdown says why. Across those 55 the semantic checks failed as `truly_absent` 29, `reads_naturally` 28, `no_collateral_loss` 18 and `spans_all_state_fact` 15 - totalling 90 across 55 pairs because a pair can fail more than one check. The partial class's 37 rejections are dominated by `reads_naturally` (29). Taking all 92 rejected pairs together, the commonest single cause is `reads_naturally` alone, then `truly_absent`; every rejected pair is itemised with its failing checks in the released dataset, and rejected pairs are **excluded rather than patched**.

The compound-fact caveat. Part of the complete class's rejection rate is a property of the targets rather than of clinical notes. A trap's text plus its correct-handling line can name several things at once ("omitting the past surgical *and* medical history"), and when the target is a compound like that, "every site of *the* fact" is ill-defined and the mapper spreads across more of the note than it should. This is the single biggest identified driver of the 44%: it is what `spans_all_state_fact` (15 failures) and `no_collateral_loss` (18 failures) are mostly catching. The cheap fix for a future build is to restrict complete-omission targets to required facts, which are single statements by

construction, and leave traps to the partial class. We did not do it: it is a scope decision taken after the build, and the affected pairs are excluded rather than reworked. What Section 3 reads off the rejection rate does not rest on those cases, because the two commonest failing checks are the other two - `truly_absent` at 29 failures, meaning the fact was still there, and `reads_naturally` at 28, meaning the note was visibly cut about - and neither is a consequence of an ill-defined target.

A.5.7 Two things the panel's second family bought

Family disagreement is concentrated and informative. The panel split on at least one field in 36 of 126 complete pairs and 21 of 151 partial pairs, most often on `truly_absent` (22 pairs), where one family read a surviving fragment as a residual and the other did not. Three seeds of a single model would have re-sampled one model's blind spots instead of surfacing them. One asymmetry is disclosed: two of the three calls come from the same family, so a lone dissent from the other family forces a re-check rather than failing a pair by itself, which is why both the majority rate and the stricter unanimous rate appear in the table above.

Construction's residual count is a property of the map, not of the note. Construction left exactly one mapped residual site on every partial pair, but the panel's own count of surviving mentions found more than one in **71 of the 151**, and its strength grading differs from construction's on the margins. A sceptic reading the errored note finds mentions the mapper did not map. Every row records both counts and both gradings, so an analysis can control for it; the distributions are per-pair fields in the released set.

A.6 The two carried-over cohorts, recorded per pair

Table A.1: Composition of the released set. What the benchmark holds, what its evaluation pairs test, and where the pairs came from. All 202 add-or-change controls are pairs carried over from an earlier build, while most omission pairs are newly built against the fact-site map, so a comparison between the two error classes is also a comparison across construction cohorts; A.6 records the four sources per pair.

The released set	
Pairs released	500
a pair is one edited note and the verified-clean twin of the consultation it came from	
Pairs in the evaluation set	495
the remaining 5 were seen by the prompt optimiser and are excluded for every judge, so all judges score an identical item set	
Consultations	117
112 of them in the evaluation set	
Verified-clean twins	112
one per evaluation consultation; the same note with the fact intact	
Matched couples: the same fact removed at two different trace levels	45
the build proposed 122 pairs naming a partner; in 45 couples both members survived verification (90 pairs)	
What the evaluation set's 495 pairs test	
Omissions	293
complete 150, fragment trace 86, restatement trace 57	
Commissions	202
additions 95, alterations 107	
Where the 500 pairs came from	
Add-or-change controls, carried over untouched	202
Built new against the fact-site map	185
Complete omissions from the earlier build	79
built by model rewrite rather than span deletion, and graded by the older single-arm severity pass	
Relabelled partial-omission seeds	34
surviving-mention counts run 1 to 9, because they predate the site map	

Table A.2: The omission half by surviving trace and severity. How much of the removed fact survives elsewhere in the note, against how much the loss matters. Severity is graded per fact on a written rubric by two models from different families, taking the lower grade where they disagree. Bracketed figures are the newly built cohort alone, which is the internally uniform one. Counts are the 293 omission pairs of the evaluation set. ‡ A restatement trace needs a fact stated twice in earnest, and peripheral facts are both the rarest and the least redundant in the corpus, so the restatement-trace peripheral cell holds only 2 pairs; its count is printed in italic and no rate is conditioned on it.

Surviving trace	Critical	Supporting	Peripheral	All
Complete	85 (20)	36 (22)	29 (29)	150 (71)
Fragment trace	44 (29)	34 (31)	8 (8)	86 (68)
Restatement trace	22 (11)	33 (33)	*2 (2)*‡	57 (46)
All	151 (60)	103 (86)	39 (39)	293 (185)

Table A.1 sets out what the released set holds, what its evaluation pairs test, and where the 500 pairs came from, and Table A.2 is the omission half by surviving trace and severity. This section records the cohorts behind Table A.1's last panel. The released set pools pairs from four sources, and two of the four were built by an earlier procedure. Rather than describe this in a footnote, the differences are fields on every pair, so any analysis can restrict to a uniform cohort. The relevant fields are `source`, `class`, `residual_level`, `severity`, `severity_source`, `residual.n_surviving` and a per-pair `provenance` block naming the builder and its inputs.

source	pairs	how built	severity grades
newly built graded pairs	185	programmatic span deletion off the site map, verified by the cross-family panel	two cross-family rubric graders with the conservative tie-break, all 185
earlier-build complete omissions	79	model rewrite under the earlier injection prompts, not span deletion, and with no site map available	66 inherited trap importance grades, 13 rubric grades from the earlier freeze
relabelled partial-omission seeds	34	earlier-build omission attempts that failed the whole-note absence check and are verified partial omissions, carried over from the earlier build	24 inherited trap importance grades, 10 rubric consensus grades
commission controls	202	earlier build, carried over untouched	inherited grades

Two consequences follow, and both are visible per pair.

The 79 earlier-build complete omissions carry older severity grades: built by model rewrite rather than span deletion off the map, graded mostly on the older single-grader path. They satisfy `omit-complete` by outcome rather than construction - each passed the same whole-note absence clause, so they are complete omissions of facts that lived in exactly one place.

The 34 relabelled partials vary in how much trace survives: they predate the site map, so the panel's count of surviving mentions runs from 1 to 9 against the newly built partials' uniform 1. Their residual strength is graded on the same three-level scale, so the trace level is comparable even though the count is not.

The newly built column is the internally uniform one and the pooled column is the bigger one, and both are reported wherever the distinction can matter. Section 3 states the one consequence that changes how a reader uses the headline result: all 202 commission controls are carried-over pairs while most omission pairs are newly built, so the headline asymmetry compares across construction cohorts, and the robustness slice restricted to the uniformly built omission pairs is what bounds it. Restricted to the 185 newly built omission pairs, the best of the eight judge designs reads 0.610

paired on omissions against the pooled 0.634, and the commission-minus-omission gap widens from 0.305 to 0.329. Every omission figure falls a little, because the carried-over cohort is slightly easier than the newly built one, and the per-fact rule of Section 7.3 reads 16.7% on the 90 newly built held-out omission pairs against 20.6% pooled, with the false-alarm cell unchanged because clean notes carry no cohort.

B THE THREE PROMPT-OPTIMISATION CAMPAIGNS

Section 6 reports that letting an optimiser rewrite a judge's prompt relocates the judge's operating point without creating the missing omission signal. This appendix carries the record behind that sentence at the depth the null needs: the three campaigns' designs and data, the held-out discipline of the confirmatory campaign, its deviations from the optimiser's published procedure, and - in B.5 - the confirmatory seed and winner printed verbatim. The candidate populations, acceptance tables and lineage records ship in the released repository (`gepa/lineage.jsonl`, `gepa/v2_results.json`, `gepa/v3_lineage.jsonl`, `gepa/v3_results.json`), each record carrying the candidate's full instruction text; the acceptance-against-held-out comparison for both confirmatory targets is Table B.1, in B.4 below.

The optimiser is reflective prompt evolution, published as GEPA (Agrawal et al., 2026). It works in a loop: run the judge prompt over a sample of training examples, hand a second model natural-language traces of what the judge got wrong, let that model propose a rewritten prompt, score the rewrite, and keep it if it scores better. The prompt being rewritten belongs to the *student*, which here is always the judge under test; the model that reads the failures and writes the rewrite is the *reflector*. Both are pinned in Appendix D. The student ran at temperature 1.0, no reasoning effort and the 1,024-token answer cap throughout, matching the settings the judge designs of Section 4 used; the reflector ran at high reasoning effort in the first campaign and medium in the third, with a 16,000-token budget. That split carries interpretive weight, because the student's budget is the one setting under which every candidate of every campaign was scored, and one of the winners behaves differently outside it: B.7 records the re-run.

B.1 WHAT EVOLVES, AND WHAT STAYS FIXED

Only one block of text changes. Everything around it - how the transcript and the note are laid into the prompt, and the format the answer must come back in - is fixed by the harness and identical for every candidate within a campaign, so a comparison between two candidates is a comparison between two instruction blocks and nothing else. The prompts of B.5 are those instruction blocks. Three wrappers are in play across the three campaigns; all three ship as release files named in B.5, so a reader can reconstruct any judge call exactly.

B.2 THE EXPLORATORY CAMPAIGN

The first campaign optimised a whole-note judge against a training pool built for it: **79 judgements over 22 consultations** - 16 additions, 22 alterations, 22 clean notes, 14 complete omissions and 5 partial omissions, 50 of them from PriMock57 and 29 from our authored scenarios. The pool shares no consultation with the 112 evaluation consultations; the overlap is asserted at load and recorded as zero in the run's own artefacts. No evaluation pair was judged at any point in this campaign.

It started from three seeds - the completeness-scored criterion of the eight designs (`seed_fc_score`), the strongest whole-note judge we could write by hand (`seed_engineered`), and the deployed faithfulness judge with an affirmative omission instruction added (`seed_v14_incl`) - ran **40 iterations**, and produced **21 candidates** (the 3 seeds plus 18 accepted mutations, with 22 further proposals rejected at a minibatch pre-filter), with zero unparseable judgements. The search objective mixed severity-weighted omission detection, commission detection and false alarms, with feasibility capped at 10% false alarms on the pool's 22 clean notes.

Two facts from its population carry into Section 6, and both survive any inspection of the released lineage. The first is the trade the population makes: **every candidate that reached 30% or better severity-weighted omission detection carried a false-alarm rate of at least 22.7%**, which on 22 clean notes is five of them; the frontier sits between the winner (`gepa-04`, a mutation of the hand-

written seed accepted at iteration 4) at 28.6% detection with 9.1% false alarms and the cheapest infeasible candidate at 31.4% with 22.7%. The second is the asymmetry, which is where Section 5 gets its earliest sighting: **all 21 candidates detect additions and alterations better than omissions** on the same severity-weighted measure, by margins running from 14.6 to 49.3 percentage points - the deployed judge, both deployable seeds, the winner, and every accepted and rejected mutation in between.

The winner's apparent gain was a shift of the score distribution, not better separation: against the seeds, severity-weighted omission detection rose from 11.4% to 28.6% at 2 of 22 clean notes flagged, while the gap between the mean clean-note score and the mean score on notes with omissions is 0.36 to 0.55 points on a ten-point scale for the winner and all three seeds alike, and on complete omissions the winner ranks the flawed note below its twin *least* often of the four (4 of 14). On data the optimiser had never seen - the study's evaluation material, 586 omission and 224 clean-note judgements - the property that made it feasible collapsed: **23.0% of omissions flagged (135/586) at a false-alarm rate of 21.9% (49/224)**, a gap of 0.35 standard errors, with paired discrimination on omissions of **0.549** and an area under the detection curve of 0.535, where 0.5 is a coin flip. No significance testing was run on the 79-judgement pool and none should be read into its numbers: 14 complete omissions and 22 clean notes mean one example moves a rate by five to seven percentage points.

B.3 The second campaign

The second campaign aimed directly at the strongest of the eight designs - the completeness-scored eight-sample judge - on a rebuilt pool of **176 judgements over the same 22 consultations**, 97 of them graded omission pairs. It ran 11 iterations and accepted one mutation, which at the design's own eight samples reads 0.418 paired on omissions against the baseline seed's 0.685 - below a coin flip, and below its own parent - and nothing in the campaign separates from the baseline (the widest gap in its tables is 3.45 points at p=0.45, and it belongs to a seed rather than to anything the search found). No prompt file was written, because writing one would have recorded a discovery that did not happen. The pool's 22 clean twins bound what the campaign can say: every false-alarm rate moves in steps of one twenty-second, so the supported claim is that prompt optimisation of this design did not beat the design's own baseline prompt here, not that prompt optimisation cannot help. Both comparison tables, the trajectory and the stall-rule extension are in the released record (`gepa/v2_results.json`).

B.4 The confirmatory campaign

Both earlier campaigns drew the reflection traces and the acceptance decision from a single small pool, which is the overspecialisation the optimiser's own documentation warns about, and B.2 is what it cost. The third campaign fixes that at full scale, runs the published mechanics rather than an approximation of them, and points the search at a compound system - the setting the optimiser's authors report it is strongest on - as well as at a whole-note judge. It is the campaign Section 6 rests on. It ran on 2026-08-14 in a single session, with zero unparseable judgements and no killed runs.

B.4.1 The partition is the experiment

The split was built and committed before any optimisation, and disjointness was asserted at every load: against the held-out subset, and against the 154 pair ids and 22 consultations the earlier campaigns had trained on, at both pair and consultation level. Everything is split at **consultation** level, because two pairs from one consultation share a transcript, a clean twin and a fact list, so a pair-level split would leak all three.

Table B.1: Two optimisation targets, what the acceptance set said about each and what the held-out set said. Paired discrimination on omissions is the tie-adjusted share of pairs in which a judge scores the note with the omission below its own verified-clean twin, so 0.500 is a coin flip. Detection is read at the most sensitive threshold whose false-alarm rate stays within 10%, and is always quoted beside that false-alarm rate. The acceptance set is 20 consultations, 48 omission pairs and 20 clean twins, and the optimiser accepted or rejected candidates on it; the held-out set is 47 consultations, 131 omission pairs and 47 clean twins, split at consultation level because two pairs from one consultation share a transcript, a clean twin and a fact list, committed before any optimisation and scored once, at the end. Differences are in points of paired discrimination, so +0.004 reads as +0.4. Intervals are consultation-clustered bootstraps, sign tests count only the pairs the two prompts ordered differently, and the detection comparison is an exact McNemar test on the flag decisions. Leg B searched at one sample per note and confirmed at eight, so its two blocks are the same criterion asked a different number of times, and the comparison that carries the result is winner against seed within each block. Leg A's acceptance-set gain, the one the search was accepted on, is not distinguishable from noise on the acceptance set itself (8 wins to 3, p = 0.23). The last row is the yardstick: the same held-out baseline run three times, spanning 0.531 to 0.649 paired and 7.6% to 13.0% detection, so a single run of this design carries about ±0.06 of paired noise - fifteen times leg B's held-out difference of +0.004. The held-out seed row's 0.801 paired is replicate 1 of the two-stage tier's three runs; Section 7.1 and Table K.5 quote that tier as the mean of the three, 0.786.

Prompt	paired omissions	det. at FA $\leq$ 10%	false alarms
Leg A - the per-fact check module inside the two-stage pipeline (15 iterations, 9 candidates, 8 accepted, 7 rejected; winner `v3p-01` at iteration 1)			
Acceptance set: 20 consultations, 48 omission pairs, 20 clean twins			
seed: the check prompt as shipped	0.792	33.3% (16/48)	10.0% (2/20)
winner `v3p-01`	0.844	33.3% (16/48)	5.0% (1/20)
difference	*+5.2 pp [−4.2, +16.2], 8 wins to 3, p = 0.23*	*0.0 pp, the same notes flagged*	*−5.0 pp*
Held-out set: 47 consultations, 131 omission pairs, 47 clean twins			
seed: the same shipped prompt, measured before the search	0.801 (replicate one)	16.0% (21/131)	8.5% (4/47)
winner `v3p-01`	0.779	16.8% (22/131)	8.5% (4/47)
difference	*−2.3 pp [−7.4, +2.7], 8 wins to 13, p = 0.38*	*+0.8 pp, 3 flags to 2, p = 1.0*	*0.0 pp*
Leg B - the monolithic faithfulness-and-completeness scoring judge (15 iterations, 9 candidates, 7 accepted, 8 rejected; winner `v3m-08` at iteration 14)			
Acceptance set: 20 consultations, 48 omission pairs, 20 clean twins			
first seed: the faithfulness-and-completeness scoring criterion, one sample	0.615	2.1% (1/48)	5.0% (1/20)
second seed: the hand-engineered omission prompt, one sample	0.646	18.8% (9/48)	5.0% (1/20)
winner `v3m-08`, one sample	0.688	25.0% (12/48)	5.0% (1/20)
difference against the first seed	*+7.3 pp [−7.0, +25.0], 13 wins to 10, p = 0.68*	*+22.9 pp, 11 flags to 0, p < 0.001*	*0.0 pp*
Held-out set: 47 consultations, 131 omission pairs, 47 clean twins			
first seed at eight samples, run 1 of 3, measured before the search	0.622	13.0% (17/131)	6.4% (3/47)
winner `v3m-08` at eight samples	0.626	8.4% (11/131)	6.4% (3/47)
difference	*+0.4 pp [−12.6, +13.3], 40 wins to 36, p = 0.73*	*−4.6 pp, 2 flags to 8, p = 0.11*	*0.0 pp*
yardstick: the same held-out baseline, its three runs	*0.622 / 0.649 / 0.531*	*13.0% / 10.7% / 7.6% (17, 14, 10 of 131)*	*6.4% / 8.5% / 6.4% (3, 4, 3 of 47)*

split	consultations	pairs	notes	complete / partial-weak / partial-strong omissions	additions / alterations	role
held-out test	47	151	198	62 / 36 / 33	10 / 10	Section 7's confirmation subset. Touched once per target, at the end
reflect	45	180	225	55 / 31 / 14	37 / 43	minibatches and reflection traces; the reflector reads nothing else
accept	20	85	105	27 / 13 / 8	17 / 20	acceptance, per-example rewards, choice of winner; the reflector never sees it

The held-out baselines were already on disk before the campaign began, bought in the runs of Sections 5 and 7, so the final comparison cost one run of the candidate and nothing else. The analysis path here recomputes the two-stage pipeline's held-out figures from those stored records and reproduces them exactly. Those figures are replicate one throughout, which is the baseline run the campaign's winners were measured against: 0.801 paired on omissions (Section 7.1's published 0.786 is the mean of three primary-family runs spanning 0.771 to 0.801), 0.935 on complete removals, 0.792 on a fragment trace, 0.561 on a restatement, and 16.0% detection (21/131) at 8.5% false alarms (4/47).

B.4.2 The mechanics that were run

The published procedure was followed on the points that matter to whether a search can overfit its own validator: candidates are kept on a per-example Pareto frontier rather than by a single aggregate score, parents are sampled in proportion to the examples on which they lead, acceptance is promiscuous (a child enters on a minibatch improvement and the frontier prunes afterwards), merge is enabled once six candidates exist, feedback to the reflector is textual and per-failure, and minibatches are drawn consultation-first so every flawed note's own clean twin sits in the same batch. The acceptance objective's primary term is paired discrimination on omissions, with detection at a false-alarm rate of 10% or below as secondary - paired is primary because the acceptance split holds 20 clean twins, so its false-alarm axis moves in five-point steps, too coarse to steer a search with.

B.4.3 Target one: the per-fact checker inside the two-stage pipeline

This is the one place a gain looked most plausible, because the pipeline of Section 7 is where the study's strongest omission signal sits and its check stage asks a small, well-posed question. The student was that pipeline with exactly one paragraph mutable. Extraction stayed frozen and cached, so the held-out run scored against the identical fact lists the baseline had been scored on, and the check stage ran at no reasoning effort through the pipeline's own code path. The search ran 15 iterations: 8 candidates accepted (7 mutations and 1 merge), 7 rejected at the minibatch, 3 merge invocations. The winner, `v3p-01`, appeared at **iteration 1**, and nothing later beat it.

Table B.1 carries the comparison. On the acceptance split the mutation gains 5.2 points of paired discrimination and halves the false-alarm rate at identical detection. On the 131 held-out omissions

it is **2.3 points worse** (cluster bootstrap over the 47 consultations [-0.074, +0.027]; sign test on the 21 discordant notes, 8 its way against 13, p=0.38) and **0.8 points better on detection**, which is one extra omission out of 131 (exact McNemar on the flag decisions, 3 against 2, p=1.0). The interval's upper end is what bounds the result: the held-out data excludes gains larger than about 2.7 points of paired discrimination. Neither prompt's detection separates from its own false-alarm rate - 1.38 standard errors for the candidate and 1.27 for the baseline - which is Section 7's picture unchanged.

The detail worth carrying is that **the acceptance-split gain was not significant on the acceptance split either**: sign test 8 wins against 3, p=0.23, cluster bootstrap [-4.2, +16.2] points on 48 omissions and 20 clean twins. The held-out set did not overturn a real effect. It declined to confirm a difference that was never distinguishable from noise on the sample used to accept it.

No prompt file was written. The rule set before the run was that a prompt file is emitted only when the winner beats its baseline on the primary measure on held-out data, and it does not.

B.4.4 Target two: the completeness-scored whole-note judge

The second target was the same judge the second campaign aimed at, at one sample during the search with the held-out confirmation at eight. Two seeds: the completeness-scored criterion, which the harness rebuilds byte for byte from the design's own prompt file, and the hand-written judge's content re-expressed in that design's frame. 15 iterations: 7 candidates accepted (4 mutations and 3 merges), 8 rejected at the minibatch, 3 merge invocations. The winner, `v3m-08`, appeared at **iteration 14** and is itself a merge of two descendants of the hand-written seed.

The nine-candidate population for this leg is likewise in the released lineage (`legs.mono.valid.all_candidates`); the winner, `v3m-08`, appeared at iteration 14 as a merge of two descendants of the hand-written seed.

A decision rule fixed before the run said to buy the held-out confirmation only if the winner beat the completeness-scored seed on the acceptance split by more than two points on both measures. It fired: **+7.3 points paired (0.688 against 0.615) and +22.9 points of detection (25.0%, 12/48, against 2.1%, 1/48)**, and it clears the hand-written seed too, by 4.2 points paired and 6.2 points of detection. So the confirmation was bought.

Table B.1 carries the held-out result beside the baseline's own three replicates on the same pairs. Against the matched single replicate the winner is **+0.004 paired** (sign test 40 against 36 over 76 informative pairs, p=0.73; cluster bootstrap [-0.126, +0.133]) and **4.6 points worse on detection** (exact McNemar 2 against 8, p=0.11). The yardstick is the baseline's own spread on these same 131 pairs: paired 0.531 to 0.649 and detection 7.6% to 13.0%, so a single run of this architecture carries roughly ś0.06 of paired noise here, about fifteen times the difference the search produced.

B.4.5 What the search wrote

Both targets produced longer, more specific, entirely sensible prompts. The first target's seed is 564 characters and its winner 3,183, with the most elaborate candidate in the population at 15,117; the winner names the clinically decisive part of a fact, polarity, certainty, dose and attribution, and tells the checker to preserve hedging. It is good advice. What it bought is visible in the held-out score distributions: mean score on clean notes 0.9587 against the baseline's 0.9612, mean score on notes with omissions 0.9315 against 0.9323, so the gap between a clean note and a flawed one is **0.0272 against 0.0289**. The prompt moved the whole distribution down about a quarter of a percentage point and left the separation where it found it - the same operating-point shift the levers of Section 6 produce, one level down, at the per-fact lookup rather than the note-level judgement.

One aside: the first target's winner has better commission calibration held out than the baseline (0.725 against 0.650, and 0.700 against 0.600 on the 10 addition pairs) - not a finding at n=20, but it points where Section 7's fabrication diagnostic pointed, at an enumerating judge's fabrication blindness being reachable by wording in a way its omission blindness is not.

B.4.6 Departures from the optimiser's published procedure

Four, all deliberate and all recorded in the run's artefacts.

1. **Merge is implemented as a reflector-mediated crossover** of two frontier candidates that lead on different examples, rather than the published module swap. Both targets evolve exactly one module - the first freezes extraction on purpose - so a module swap has nothing to swap between.
2. **The second target's hand-written seed is not character-for-character identical** to the engineered judge (Table K.2) that shares its content: it is that prompt's content re-expressed in the frame the eight designs use, with the transcript and note moved to the front. The completeness-scored seed *is* identical, character for character, to that design's own prompt.
3. **The acceptance objective's primary term is paired discrimination rather than single-note detection**, for the sample-size reason stated with the mechanics above.
4. **The search ran at one sample per judgement on both targets**; only the second target's held-out confirmation ran at eight.

The two earlier campaigns departed further, and in ways the third was built to fix: both drew reflection traces and acceptance decisions from one pool, the first accepted candidates on an 8-example unstratified minibatch rather than full-pool scores, and the second searched at four samples with one seed left unconfirmed at eight and extended its own stall rule mid-run. Every departure is recorded in the released campaign records.

B.5 The confirmatory seed and winner, verbatim

B.4 records the confirmatory campaign growing a 564-character seed into 3,183 characters of clinically sensible instruction that bought nothing on held-out data. That pair is reproduced verbatim below, exactly as stored in the campaign's lineage record, so the reader can see the object the claim is about. **Every other seed, winner and harness wrapper of all three campaigns ships verbatim in the released repository** - the campaign lineage files carry every candidate's full instruction text with its hash - and the inventory below identifies each one so a reader can pull and verify any of them byte for byte: the inventory - every seed, winner and wrapper of all three campaigns, with character counts and SHA-256 prefixes - is in the released lineage records named at the top of this appendix, and the three harness wrappers are the release files `gepa/gepa_data.py`, `w2_prompts/grid_FC_score.txt` and `w2_prompts/pipeline_check_binary.txt`, so any judge call in any campaign can be reconstructed exactly.

B.5.1 The confirmatory seed, verbatim

The check-stage paragraph as shipped, 564 characters:

```
For each reference fact, decide whether the clinical note CAPTURES it -
    stated verbatim, stated
in different words, or unambiguously implied by what the note does say.
    A faithful paraphrase, a
correct unit or dose conversion, an accepted abbreviation, and the same
    fact recorded in an
unexpected section of the note all count as captured. Anything the note
    simply does not carry, in
any form, is not captured.

Judge only whether each fact is there. Do not judge the note's style,
    length, ordering or
formatting, and do not penalise it for carrying extra material.
```

B.5.2 The confirmatory winner, verbatim

`v3p-01`, the mutation that gained 5.2 points on the acceptance split and was 2.3 points worse held out, 3,183 characters:

```
For each numbered reference fact, judge whether the note conveys the
    same clinical proposition, not merely whether it contains related
    words.
```

```
First identify the fact's load-bearing content: the
    patient/finding/action involved; polarity or negation;
    timing/currentness; site/laterality; value, dose, route or frequency
    if relevant; who said or did it when attribution matters; certainty
    level such as possible/probable/confirmed/ruled out/not established;
    and any stated reason, red flag, pertinent negative, diagnosis,
    plan, urgency, or "only/no other" qualifier that changes the
    clinical meaning.

Mark a fact present when the note carries that load-bearing content
    explicitly, by faithful paraphrase, by accepted abbreviation, by a
    conventional clinical phrase, or by an unavoidable implication from
    a heading or context. It may be in any section and need only be
    recorded once. Do not require conversational wording, duplicate
    mentions, exact order, or unnecessary detail. A broad but clinically
    standard statement can cover a narrower fact when its scope clearly
    includes it, such as a relevant normal examination phrase, a
    system-wide negative, or a medication/allergy/status entry under the
    appropriate heading.

Mark a fact absent when the note does not let a reader recover the
    load-bearing proposition from the note alone. In particular:
- A mere mention of the same topic, symptom, medication, diagnosis, or
    plan is not enough if the required relationship is missing.
- A surviving fragment is not enough if the fact depends on the dropped
    element, such as the working diagnosis, explanatory link, reason for
    urgency, dose, current medication status, "no other" qualifier, red
    flag, or pertinent negative.
- If the note states an incompatible version of the fact, mark the
    reference fact absent: opposite polarity; present vs denied;
    assessed vs not assessed; different dose/value/timing/site; wrong
    source or attribution where source matters; or a symptom term used
    in a different clinical context.
- Preserve certainty. A possible, suspected, differential, working, or
    to-be-excluded diagnosis is not captured by a note that records it
    as confirmed; a confirmed diagnosis is not captured by a note that
    records only a query; "not established" or "not ruled out" is not
    captured by documenting the finding as absent.
- For medication, contraception, allergy, past history, social history,
    and occupation facts, require the status the fact asserts.
    Mentioning an item somewhere else does not by itself establish that
    it is current, the only medication, absent, denied, or part of the
    formal medication/allergy/history list unless the note's wording or
    section unambiguously does so.
- For patient ideas, concerns, expectations, clinician explanations,
    counselling, safety-netting, and plans, require the substantive
    content being attributed, not just that discussion occurred.

Be generous about faithful clinical compression, abbreviations, and
    section placement, but do not fill in missing qualifiers from
    medical common sense or from what would usually be true. The
    question is whether this note actually carries this fact, with its
    clinically important meaning intact.
```

B.6 THE PORTABLE LESSON, AND THREE COROLLARIES

Across two targets, 30 iterations and 15 accepted candidates, the confirmatory campaign moved the held-out estimate by -2.3 and +0.4 points of paired discrimination, neither distinguishable from zero, against a baseline whose own run-to-run spread on the same pairs is 0.531 to 0.649. The narrow claim is the one Section 6 makes: prompt optimisation, run properly, did not beat the prompts we already had, on either a whole-note judge or the per-fact checker inside the pipeline.

The methods result is more portable, and it is the reason this appendix prints the acceptance tables next to the held-out ones. **A 20-consultation acceptance split will hand you a five-to-seven-point improvement that a 47-consultation held-out set does not see.** Both the first and the third campaign found that; the first found it only after emitting a prompt file, and that prompt then went on to read 21.9% false alarms on data it had never seen. Three corollaries follow, and they are cheap to apply.

Split at the unit that shares structure, not at the row. Two pairs from one consultation share a transcript, a clean twin and a fact list, so a pair-level split leaks all three and the validator stops being independent of the training signal.

Measure the held-out baseline before the search starts. The held-out figures for both baselines here were already bought and on disk before the campaign began, so the final comparison cost one run of the candidate and the held-out set was touched exactly once, at the end.

Emit the artefact only on the held-out result. The third campaign's first target was chosen on a 5.2-point acceptance-split gain that its own sample could not resolve - 8 wins against 3, p=0.23 - so a rule keyed to the acceptance split would have shipped it, and the held-out set says there was nothing to ship. The rule used instead - a prompt file is written only when the winner beats its baseline on the primary measure on held-out data - meant no file was written for either target. The only prompt file any of the three campaigns emitted is the first campaign's, written under the loose rule, and B.2 is what happened to it.

B.7 The budget every candidate was scored in, and what happens when it is lifted

One setting is common to all three campaigns and appears in none of their result tables. Every candidate prompt was executed by the student with no reasoning effort inside the 1,024-token answer cap of Section 4 - `gepa_optimize.py`, `gepa_v2.py` and `gepa_v3.py` each pass `reasoning_effort="none"` at their judging call - while only the reflector, which reads the failure traces and writes the rewrite, was given room to deliberate. A candidate could therefore be proposed with an instruction it had no budget to carry out, and then be scored as though the instruction had been tried.

That is what happened to the exploratory campaign's winner. `gepa-04`'s instruction block leads with an enumerate-then-check procedure, and B.2 records it reading 23.0% of omissions flagged at 21.9% false alarms on data it had never seen, paired 0.549. Re-run unchanged, character for character - the same released prompt file, `gepa/judge_prompt.txt`, SHA-256 prefix `bb47c133` - on the 151-pair held-out subset at high reasoning effort with a 12,000-token cap, three seeds, it reads **0.669 paired against 0.546 for its own effort-none control on that same subset** (change +0.123, consultation-clustered 95% interval [+0.059, +0.189]), and its clean-note score distribution changes shape: 45, 46 and 45 of the 47 clean twins land on exactly 10 out of 10, which is the calibration behind the evolved prompt's held-out operating point, 32.1% detection at 2.1% false alarms. The prompt text is identical character for character across the two runs; only the reasoning budget it executed under differs.

The same re-run carries the control that keeps this from being a claim about compute in general. The engineered judge - the strongest hand-written judge in the study - re-run the same way, gains **+0.013 [-0.047, +0.071]** - an interval straddling zero on a prompt already at 0.635 - while the plain completeness-scored criterion gains **+0.101 [+0.048, +0.153]**. A reasoning budget is worth about a tenth of a point of discrimination to some prompts and nothing to others; Table B.2 carries the full ladder with the compute-matched comparison against the pipeline, and Section 7 the reading. Every prompt text involved is unchanged from B.5 and the released inventory, and the re-run's records, per-replicate tables and clustered intervals ship in the data repository under `judgements/judges/w2-power/`, beside the study's other judge families.

Two things this does not establish. It is one already-selected prompt given a budget, not a search conducted at one: a search with a larger student budget would make different acceptance decisions over a different candidate population, and it was not run. And the clean-note calibration is a behaviour observed on three seeds on this corpus, not a property the prompt guarantees, so a deployment would have to re-establish the threshold on its own clean notes (Appendix L).

Table B.2: What a reasoning budget buys each prompt, on the 151-pair held-out subset. One set of pairs throughout: 131 omission pairs and 20 commission pairs over 47 consultations, with 47 clean twins. Within each block the prompt is identical, character for character, between the control row and the rows beneath it, and the judge model, temperature, seed, notes and flag rule are held, so the reasoning setting and the token cap are the only quantities that move. Paired discrimination is the tie-adjusted probability that a note carrying an omission scores below its own verified-clean twin, where 0.500 is a coin flip; a value here is the mean over the runs in the "runs" column. Intervals are consultation-clustered bootstrap 95% intervals on the change, 10,000 resamples. Four rows rest on a single run and should be read against their control's own run-to-run spread, which on this subset is 0.508 to 0.603 for the completeness-scored single-call design. Reading: the budget is worth about a tenth of a point of discrimination to a plain criterion and to the optimiser's winner, nothing to the engineered omission judge, and the strongest monolithic judge it can build, with the fact list and the budget together, still sits about 0.09 below the pipeline reading a list of the same kind through closed per-fact verdicts - the audited 37.6-fact list in the prompt against the two-stage tier's own 39.2 unaudited facts (0.786 as a mean over three runs spanning 0.771 to 0.801).

Prompt (unchanged within each block)	reasoning budget	runs	paired, omissions	change against its own control
Faithfulness and completeness, score, one call	none, 1,024	3	0.570	reference
	medium, 12,000	1	0.641	+0.071 [+0.009, +0.134]
	high, 12,000	3	0.670	+0.101 [+0.048, +0.153]
The same, with the pipeline's audited fact list in the prompt	none, 1,024	1	0.634	reference
	high, 12,000	1	0.699	+0.065 [−0.016, +0.145]
The engineered omission judge	none, 1,024	2	0.635	reference
	high, 12,000	1	0.649	+0.013 [−0.047, +0.071]
The optimiser's exploratory winner	none, 1,024	2	0.546	reference
	high, 12,000	3	0.669	+0.123 [+0.059, +0.189]
Two-stage pipeline, closed per-fact verdicts	one call per fact, none	3	0.786	not a monolithic judge

B.8 THE EXPLORATORY WINNER, VERBATIM

The prompt behind Section 7's evolved prompt - `gepa-04`, the exploratory campaign's winner, released as `gepa/judge_prompt.txt` (SHA-256 prefix `bb47c133`, 9,071 characters) - is reproduced in full below, exactly as stored, since it is one of the two objects the paper's central result rests on. Its instruction block opens with the enumerate-then-check procedure B.7 discusses; everything else is what the reflector added around that instruction over the campaign's forty iterations, starting from the hand-written seed B.2 names. The braces mark the two slots the harness fills per call.

```
CONSULTATION TRANSCRIPT:
{transcript}

CLINICAL NOTE (AI-generated from the transcript above):
{note}

---

You are a clinical safety reviewer assessing whether a clinical note is
    faithful to, and complete from, the consultation transcript it was
    generated from.

Your central question is: would a competent clinician reading only the
    note know the clinically load-bearing facts established in the
    consultation, without being misled by anything the note added,
    inverted, hardened, or changed?

Do not judge by how polished the note is. Judge by clinical content.
```

```
Review method: compare in both directions

1. Transcript-to-note audit: first build a mental inventory of the
    clinically important facts established in the consultation, then
    check whether the note carries them.
2. Note-to-transcript audit: check every clinically meaningful
    assertion in the note against the transcript. A wrong assertion is
    as serious as a missing one.

Clinically important facts to look for include, when present in the
    transcript:

- Presenting problem(s), key chronology, duration, progression,
    severity, location, laterality, triggers, functional impact, and
    associated symptoms.
- Pertinent positives and pertinent negatives, especially red-flag
    questions and answers.
- Symptoms that broaden the differential even if they are not the main
    complaint, such as exertional breathlessness, chest pain, syncope,
    fever, weight loss, neurological symptoms, bleeding, severe pain, or
    suicidal/self-harm thoughts.
- Examination findings, observations, abnormal findings, and important
    normal findings used to justify reassurance or conservative
    management.
- Working diagnosis, differential diagnosis, level of uncertainty, and
    the reasoning behind tests, referrals, or conservative treatment.
- Medication names, doses, routes, frequency, duration, changes,
    stopped medication, allergies, adverse reactions, contraindications,
    pregnancy status, and other safety-relevant history.
- Investigations requested, results discussed, referrals made, urgency,
    target service, reason for referral, and any timeframes.
- Follow-up and safety-netting: what to do, when to do it, which
    symptoms should trigger action, and where to seek help.
- Safeguarding, capacity, mental health risk, social circumstances,
    occupation/driving advice, or administrative outcomes when these
    affect care.

Omission types you must catch

A fact can be omitted even when the general topic is still mentioned.
    Ask what a reader would actually know from the note.

Count it as an omission if the note preserves only a vague shell but
    loses the load-bearing part, for example:

- It says "bloods requested" but not the suspected condition or
    clinical reason that makes the results actionable.
- It says "referral made" but not that it was urgent, cancer-suspected,
    same-day, or driven by red-flag features.
- It mentions "safety-net given" but omits the specific trigger, route,
    or timeframe that matters.
- It lists a symptom area but drops the severity, duration, exertional
    nature, laterality, neurological feature, bleeding, fever, weight
    loss, or other detail that changes risk.
- It records a medication but omits the dose, frequency, duration, or
    change where that would affect safe prescribing.
- It records a diagnosis but drops the clinician's uncertainty or
    differential when that uncertainty was important to management.
- It mentions general aches or symptoms elsewhere but fails to carry a
    specific red-flag answer, such as pain/stiffness with neck movement,
    exertional symptoms, waking from sleep, vomiting, neurological
    symptoms, or self-harm denial.

Altered-content types you must catch
```

```
Penalise clinically meaningful changes even if nothing is "missing" in
    the ordinary sense. In particular, actively check for:

- Flipped negation or polarity: "denies vomiting" becoming "vomiting";
    "not waking from sleep" becoming "waking from sleep"; "no suicidal
    ideation" becoming suicidal ideation; "no chest pain" becoming chest
    pain, or the reverse.
- Wrong value: incorrect dose, frequency, duration, age, temperature,
    blood pressure, date, number of episodes, timeframe, or
    investigation result.
- Wrong laterality or site: left/right, upper/lower,
    unilateral/bilateral, temporal/occipital, arm/leg, etc.
- Wrong subject: patient versus partner, child, parent, family history,
    clinician, or previous doctor.
- Hardened uncertainty: "possible", "likely", "to exclude", or
    "watchful waiting" turned into a definitive diagnosis or exclusion
    when that would affect care.
- Fabricated findings, diagnoses, allergies, examination signs, advice,
    referrals, or medications not supported by the transcript.
- Over-reassurance: "no red flags" or "normal examination" when the
    transcript contains a positive, equivocal, unasked, or undocumented
    red-flag point that matters.

Use the transcript as the source of truth. If the transcript is
    genuinely ambiguous, do not penalise a clinically reasonable
    paraphrase. But do penalise the note if it creates false certainty,
    a false denial, a false positive, or a misleadingly complete
    statement.

What not to penalise

Clinical notes legitimately compress. Do not penalise:

- Paraphrase into clinical language.
- Standard abbreviations and conventional normal-exam phrasing, if
    consistent with what happened.
- Combining repeated statements into one entry.
- Recording a fact once rather than every time it was said.
- Dropping conversational colour, pleasantries, hesitations, filler, or
    non-clinical digressions.
- Omitting exhaustive symptom detail when the load-bearing clinical
    content is still present.
- Different ordering, formatting, or note style.
- Minor wording differences in safety-netting or plans when the same
    trigger, urgency, and route remain clear.
- A concise note that a careful clinician would still regard as
    faithful and sufficient.

Severity calibration

Grade the consequence of the missing or wrong content, not the
    emotional intensity of the consultation.

Critical:
Use this when the absence or alteration could plausibly change
    diagnosis, triage, urgency, treatment, prescribing safety,
    investigation, referral, follow-up, safeguarding, or risk
    assessment; or when it removes a safety net or creates a latent
    safety risk. This includes important red flags, pertinent negatives
    used to rule out serious disease, allergies, medication
    changes/doses, abnormal examination findings, urgent referral
    reasons, escalation advice, and safety-net triggers. Any critical
    issue must receive a score low enough to be flagged for human review.
```

```
Supporting:
Use this when the issue materially weakens the clinical record,
    reasoning, handover, or defensibility, but is unlikely by itself to
    change immediate care. Examples include missing context for
    conservative management, incomplete chronology, or a non-urgent
    rationale that would help future interpretation. Supporting issues
    usually should not be flagged unless several together materially
    undermine the note.

Peripheral:
Use this for details with no meaningful bearing on diagnosis,
    management, risk, handover, or defensibility. Do not penalise
    peripheral omissions.

Scoring calibration

- A faithful and complete note should receive full credit, with no
    speculative issues added.
- Do not reduce the score merely because the note is concise or uses
    different wording.
- One critical omission, inversion, fabrication, or clinically material
    alteration is enough to make the note fail human-review threshold.
- Multiple critical problems, or a fabricated/inverted red flag,
    medication safety issue, allergy issue, or urgent-care issue, should
    receive a substantially lower score.
- Supporting issues may reduce the score modestly; multiple supporting
    issues can reduce it further if their combined effect materially
    weakens the record.
- If you are genuinely torn between supporting and peripheral, choose
    peripheral and do not penalise.
- If you are genuinely torn between critical and supporting, choose
    supporting only when patient care, safety-netting, triage,
    prescribing, or diagnostic risk would probably be unchanged.

Final check before scoring

Ask yourself:

1. What clinically important facts did the transcript establish?
2. Would a reader of the note know those facts?
3. Did the note change any yes/no answer, dose, timeframe, subject,
    laterality, diagnosis, urgency, or safety-net instruction?
4. Did a surviving vague mention hide the loss of the load-bearing
    detail?
5. Are any concerns merely stylistic compression or conversational
    detail?

Only penalise differences that matter clinically or materially weaken
    the record.

Return ONLY a JSON object on one line, with no other text before or
    after it:
{"score": <integer 0-10>, "omissions": [{"missing": "<what the note is
    missing, or what it asserts wrongly>", "severity":
    "critical|supporting|peripheral", "why": "<one line: what could go
    wrong because of it>"}]}

The score runs one way and one way only: 10 = the note fully satisfies
    the criterion, 0 = a severe violation. Leave "omissions" empty if
    you are penalising nothing.
```

C PIPELINE DIAGNOSTICS AND DEPLOYMENT PARAMETERS

This appendix carries the working detail behind Section 7: what each stage of the enumerate-then-check pipeline does and what it is allowed to see, the reasoning-effort deviation and the experiment that resolved it, two diagnostics that looked like mechanisms and did not survive the held-out run, the economics of the quote-verified second look, and the parameters a team would need to run either tier themselves. Two sets of pairs appear throughout and are never pooled. The **exploratory pass** covers 38 pairs against 32 clean notes, 70 note-judgements per tier. The **confirmation** covers 151 held-out pairs - 131 omissions (62 complete, 36 partial-weak, 33 partial-strong) plus 10 addition and 10 change pairs - against 47 clean twins over 47 consultations, 198 note-judgements per tier per replicate; both tiers carry three replicates (seeds 11, 22 and 33; extraction and audit cached across all three), with zero parse failures in any tier or replicate, and the diagnostics in this appendix read replicate 1. Pair-id overlap between the two is zero. Following Section 7, **the two-stage pipeline** is B2 in the released data and **the three-stage pipeline** is B3; the codes are the released data's names for the two tiers.

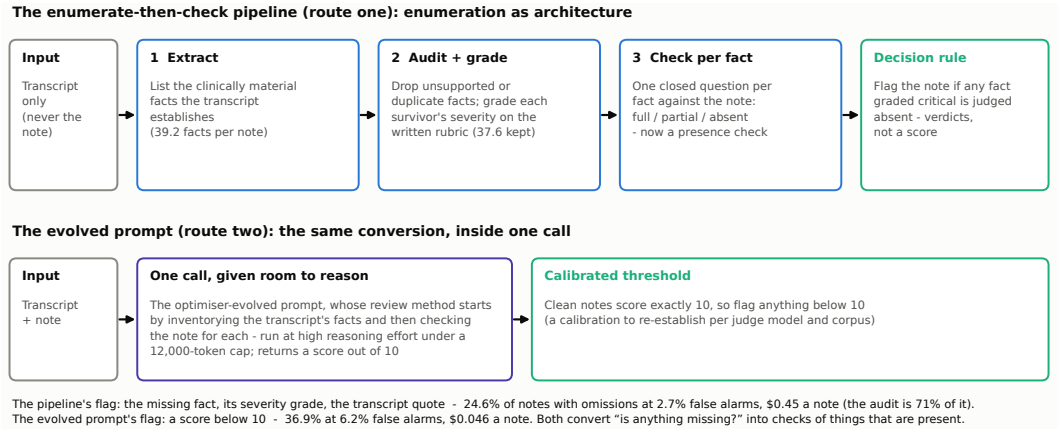


Figure C.1: The two methods, as run. The enumerate-then-check pipeline (route one in the released data, and in the figure) converts the absence question into presence checks in the architecture: an extraction call that never sees the note, an audit that drops weak facts and grades the rest, one closed check per fact, and a decision rule over the verdicts. The lane draws the three-stage tier, the only one whose audit grades severity and so the only one the decision rule can be stated on; the two-stage tier stops after the per-fact check and carries the higher paired score. The three-stage tier also runs a second look over the facts the check called absent, which the lane does not draw (Section C.1 below). The evolved prompt (route two there) makes the same conversion inside one call: a GEPA-evolved prompt whose review method starts by inventorying the transcript's facts and then checking the note for each, executed at high reasoning effort and read through a clean-note calibration. Operating points are the evaluation-set figures of Table K.5 and prices those of Table K.6; the fact counts inside the pipeline's lane are the held-out confirmation run's, 39.2 facts extracted per note and 37.6 left after the audit.

C.1 THE STAGES, AND WHAT EACH ONE IS ALLOWED TO SEE

The pipeline runs the same operations the benchmark's construction used, at inference time, in four stages - the three that name the three-stage tier, plus a second look that fires only on flagged absences. **Extraction** reads the transcript and writes a numbered list of clinically material facts, each with an id and its supporting evidence; it runs once per consultation and is cached to disk. **The audit** (three-stage pipeline only) reads the transcript and that candidate list, drops facts that are unsupported, duplicated or not something a note would be expected to record, and grades each survivor's severity against the written rubric of Section 3 (printed in Appendix A); it also runs once per consultation and is cached. **The check** is the only stage that sees the note under judgement: it receives the note and the numbered fact list and returns one verdict per fact id, present-or-absent on the two-stage pipeline and full/partial/absent on the three-stage pipeline. **The second look** (three-stage pipeline only) receives the note and only those facts the check verdicted absent, and may rescue a fact by finding it in the note.

Three contract details matter for anyone reproducing this. Verdicts are keyed to fact ids rather than returned as a list, so a mis-count is a set difference rather than a silent misalignment; ids that come back unanswered trigger one targeted retry asking only about them, and anything still

unanswered makes the record a parse failure rather than a fraction over an unknown denominator. In the confirmation neither tier needed that retry on any of its 198 note-judgements, and neither produced a parse failure. The second look must return the verbatim note text carrying any fact it claims the check missed, and a rescue is accepted only if that quote actually occurs in the note after whitespace and case normalisation, at eight characters or more; an unverifiable quote is counted and the absence stands. And the fact list shown to the check stage carries ids and fact text only: severity is withheld deliberately, because a checker told that a fact is critical is a checker being nudged about how hard to look.

Scoring and thresholds. The two-stage aggregate is the share of facts present. The three-stage aggregate is severity-weighted capture, crediting full 1.0, partial 0.5 and absent 0.0, with weights critical 4, supporting 2 and peripheral 1; those weights are an indicative guess at relative clinical cost rather than a measured quantity, and the unweighted variant is recorded beside every note so nothing rests on them. Both tiers write their flag field as null on purpose: no flag threshold was fixed before the run, and every operating point in Section 7 and below is swept in analysis and quoted with the false-alarm rate at that same threshold.

What the stages produce, on the confirmation's 47 consultations: extraction yields 39.2 facts per note, and the audit cuts that to 37.6. The three-stage pipeline's 198 note-judgements therefore rest on 7,443 per-fact verdicts, 6,548 full, 675 partial and 220 absent after the second look; the two-stage pipeline's rest on 7,762 binary verdicts, 7,291 present and 471 absent.

C.2 Blindness is a structural assertion, not a convention

Extraction and audit must never see the note under judgement, because a fact list contaminated by the note is a fact list built from the thing it is used to score. Two mechanisms enforce that. The prompt builders for both stages take a transcript, and for the audit a candidate fact list, and nothing else; no note text is in scope inside either function. Before either call is made, `assert_blind` compares the built prompt against every note of that consultation: it normalises whitespace and case, takes three 120-character probes from each note (start, middle, end), and raises rather than running if any probe occurs in the prompt. A note is a generated structured document rather than a span of the transcript, so a hit would be a real leak instead of incidental overlap. Because the assertion raises rather than degrading, a completed run is itself the record that it never fired: all 37 extraction and 37 audit calls of the confirmation completed.

One further separation follows from the pinned roles rather than from an assertion: the model that writes the fact list is from a different family from the model that checks the note against it (Appendix D), so the enumerated facts are not the checking family's own.

C.3 The reasoning-effort confound, and the experiment that removed it

The eight monolithic judge designs ran their judge at no reasoning effort. The pipeline's check stage did not, and the reason is recorded in the runner. A two-consultation wiring pilot (4 note-judgements per tier) found that at no reasoning effort the checker answered a 36-fact keyed question in about 190 output tokens with zero reasoning tokens, roughly five tokens per verdict, and that across the pilot's four three-stage note-judgements it never once used the trichotomy's middle label: 147 full and 3 absent, no partials. A four-call re-probe at medium effort on the two clean notes of that pilot, costing $0.06 and stored nowhere, produced 1,100 to 1,200 reasoning tokens and did use "partial". The check stage was therefore moved to medium effort, which left the confound plainly on the table: any win over the monolithic designs might be deliberation rather than task restructuring.

We tested it directly by re-running the two-stage pipeline on the identical 38-pair exploratory subset at no reasoning effort, changing nothing else. Effort was not the source of the gain. At no effort the tier scores 0.828 paired discrimination on omissions, where 0.5 is a coin flip, with swept detection of 31.2% at a 9.4% false-alarm rate; at medium effort on the same pairs it scores 0.813 paired and 25.0% detection at the same 9.4% false alarms. It is also cheaper, $0.0074 against $0.0346 per note-judgement, a factor of 4.65. The confirmation therefore ran the two-stage pipeline at no reasoning effort, the same setting the monolithic designs used, which removes the confound for that tier outright; the three-stage pipeline kept medium effort because its trichotomy demonstrably needs it. The records agree: all 198 two-stage confirmation note-judgements carry zero reasoning tokens,

while the three-stage pipeline averages 1,834 reasoning tokens per note-judgement across its check and second-look calls combined.

C.4 THE TRICHOTOMY DIAGNOSTIC DID NOT REPLICATE

The analysis script re-scores the same purchased verdicts under variants. The most informative of them, "lenient", credits a partial verdict as full, so only an outright absent counts against a note; it is the cleanest read available on whether the trichotomy's middle label is doing work. On the exploratory pass it reached **46.9% detection at 6.2% false alarms** (32 omission pairs, 32 clean notes), the strongest number that pass produced. On the 131 held-out omission pairs against 47 clean twins it reaches **22.9% at 8.5%**. It more than halved. These variants are attribution diagnostics computed on verdicts already bought, not a menu of configurations to choose from, and picking the best of them would have been tuning on the exploratory pass, which is exactly what the halving punishes.

The trichotomy also failed the other half of what it was expected to do. It was designed on the expectation that the largest gain would fall on partial omissions, where the removed fact leaves some trace behind. At a swept threshold holding false alarms at or below 10%, the three-stage pipeline detects 19.4% of the 62 complete omissions, 17.4% of the 69 partial omissions (weak and strong pooled) and 12.1% of the 33 partial-strong ones. On the paired measure the ordering is the same and wider: complete 0.863 against 0.652 for partials pooled, and within partials 0.819 weak against 0.470 strong. What the audit stage actually bought was the severity grades, which are the ingredient the per-fact decision rule of Section 7 needs and the cheaper tier cannot supply: all 471 of the two-stage pipeline's missing-fact entries carry no severity at all.

C.5 THE QUOTE-VERIFIED SECOND LOOK: REAL WORK ON FACTS, NO CHANGE AT NOTE LEVEL

On the confirmation, the second look examined **299 flagged absences and rescued 79 of them, 26.4%**, with **zero** rejections for an unverifiable quote; the quote-verification rail was live throughout and never had to fire, and no second-look call errored. Its effect on the headline is a wash. Re-scoring the same verdicts without it gives 0.752 paired discrimination on omissions with the stage against 0.763 without, and swept detection of 18.3% against 17.6% on the 131 held-out omission pairs.

It is not free. The stage fires only on note-judgements carrying at least one flagged absence, which was 124 of the confirmation's 198, so it adds 0.63 calls per note-judgement rather than one. Those 124 calls cost $2.77 of the $8.71 the three-stage pipeline spent on note-level calls across the confirmation's 198 note-judgements, which is **$0.014 of that pipeline's $0.044 marginal cost per note, or 32%**, for no measurable change in either direction. The 26.4% flip rate says the second look is doing real work on individual facts; the note-level figures say that work cancels out. Section 7 keeps the stage inside the three-stage tier, whose audit grades are what the per-fact rule fires on; a deployment that does not need the rescue saves 32% of that tier's per-note spend by dropping it.

C.6 FABRICATION ANTI-CORRELATION ON THE ADDITION PAIRS, AT N=10

The confirmation's 10 addition pairs - a fabricated fact inserted into an otherwise clean note - give the three-stage pipeline a paired score of **0.300**, below the 0.5 of a coin flip, which means it scores the note containing the fabrication above its own clean twin more often than not. The two-stage pipeline on the same 10 pairs scores 0.600, the three-stage pipeline on the 10 change pairs 0.700, and the exploratory pass's 3 addition pairs gave the three-stage pipeline 0.833. The mechanism is plausible enough on its face: a coverage score rewards a note that says more, and neither tier ever asks whether the note contains content the transcript does not support, which is the mirror-image blindness Section 7 describes. At 10 pairs this is a signal to size properly rather than a finding, and it is reported that way in the main text.

C.7 DEPLOYMENT PARAMETERS

The full deployment parameterisation - per-stage model roles, reasoning-effort settings, what each stage receives, call schedules, caching keys, and the measured per-note costs in both framings - ships with the release rather than here: the pipeline runner (`w2_pipeline.py`), its prompt files (`w2_prompts/`), and the run manifests carry every setting, and Appendix D records the per-note prices with their measurement dates. Three facts from that record stay in the paper because results rest on them. At the one-note-per-consultation price, the audit stage is 71% of the three-stage pipeline's $0.45 per note, which makes folding severity grading into the extraction call, and a cheaper extraction model, the two untested cost levers (Appendix L). The confirmation bought 37 extractions and 37 audits for its 47 consultations, because 10 of those consultations also appear in the exploratory pass and their cached stages were reused - so the confirmation is a fully held-out test of the judging and only a partial one of the extraction, as Appendix L states. And latency was not measured: the harness records token receipts, cost and a completion timestamp per call, but no per-call wall-clock duration, so no latency figure exists and none should be inferred from the timestamps.

D REPRODUCIBILITY

This appendix records what a replicator needs in order to regenerate the study: which model ran in which role, on what transport and at which settings, and how runs were seeded, hashed and identified. The reproducibility statement lists the harness, prompts, judge configurations, model pins and run manifests among the released artefacts, and this appendix is the description of those objects. The per-note prices a practitioner would pay are in Table K.6, and are reconciled to their measurement dates in D.3 below.

The full versions of what this appendix summarises are release material rather than print. The lock file itself (`models.lock.json`) and the validator every manifest was written to satisfy (`check_manifests.py`) ship with the harness; the run manifests themselves, with their per-call logs and the prompt and dataset digests every manifest records, ship in the data repository beside the judgements they describe. Transports, hashes and the manifest schema are therefore checkable in the release rather than printed here. Of the run manifests, those of the released benchmark judge runs ship, complete runs only; incomplete runs are excluded at packaging. Each manifest also records the state of the working tree its run started from, including the exact uncommitted paths where there were any.

D.1 MODEL PINS

Every call that stands behind a number in this paper goes through a single function that resolves a named role against a lock file, `models.lock.json`, and refuses to run if the role is not pinned there. The lock records, per role, the provider slug, the provider-routing block sent with each request, and the credit price per million tokens at pin time. It was pinned on 2026-07-29, and extended twice while the study ran: the construction role gained an explicit routed entry on 2026-08-10 when construction moved off a subscription path onto metered credits, and the second-family judge was added on 2026-08-17. Slugs and prices were verified live against the provider's model listing on each of those three dates.

The manifest counts in the table were recomputed on 2026-08-17 from the 424 run manifests then present; at the 25 Aug freeze the tree holds 569 run manifests, 539 of them complete, and the 240 describing the released benchmark judge runs ship in the data repository, alongside the census pipeline's own manifests. A run that used two roles appears in both rows. Two rows carry pins that were used lightly and are reported as such rather than as further judge families: the open-weights row appears in one run, and the second Anthropic row in nine. The Anthropic route lists no seed support, so seeding there is best-effort and the run-to-run spread, not the seed, is what stands in for determinism.

The second-family judge and its reasoning setting. The eight ablation designs run their judge at no reasoning effort. The second family's endpoint refuses that setting outright, returning an HTTP 400 with "Reasoning is mandatory for this endpoint and cannot be disabled" when probed on 2026-08-17, so "minimal", the nearest available analogue, is what is pinned. Judge family and reasoning setting therefore move together in every second-family comparison, and no run in this study can

role (lock key)	pinned model	transport and route	pinned settings	credit price in / out per Mtok	where it ran
`judge-primary`	`openai/gpt-5.4`	OpenRouter, provider order `["openai"]`, fallbacks off	temperature 1.0, reasoning effort none, 1,024 max output tokens; three disclosed deviations, each recorded per run (the three-stage pipeline's check stage at medium effort, the census verification panel at high, and the cached per-consultation extractions and checklists at temperature 0)	$2.50 / $15.00	the judge in all eight ablation designs, every reference judge, the RAGAS-style recipe, the external commission anchor, the check stage of both pipeline tiers in the primary family, and the student prompt in all three optimisation campaigns (188 run manifests)
`auditor`	`openai/gpt-5.5`	OpenRouter, `["openai"]`, fallbacks off	temperature 1.0, reasoning effort high, except medium in the census work and as the reflection model of the confirmatory optimisation campaign	$5.00 / $30.00	the answer-key audit and its semantic checks, the omission-pair verification panel, the severity grading, the pipeline's audit stage and its quote-verified second look, and the reflection model in the optimisation campaigns (175 manifests)
`constructor`	`anthropic/claude-opus-5`	OpenRouter, `["anthropic"]`, fallbacks off	temperature 1.0, reasoning effort medium, 16,000 max output tokens	$5.00 / $25.00	benchmark construction (reference-note repairs, regenerations, severity grades, omission injection, the fact-site map), the census discovery passes and cluster labelling, and the pipeline's extraction stage (112 manifests)
`judge-gemini`	`google/gemini-3.1-pro-preview`	OpenRouter, `["google-vertex"]`, fallbacks off	temperature 1.0, reasoning effort minimal, 1,024 max output tokens	$2.00 / $12.00	the second-family replication of two designs, and the check stage of both pipeline tiers in the transfer run (24 manifests)
`judge-opus`	`anthropic/claude-opus-4.8`	OpenRouter, `["anthropic"]`, fallbacks off	temperature 1.0, 1,024 max output tokens; high reasoning effort in the verification panel, none sent in the second-family probe	$5.00 / $25.00	the second member of the cross-family omission-verification panel, and the second-family candidate probe (9 manifests)
`judge-qwen`	`qwen/qwen3-235b-a22b-2507`	OpenRouter, no provider constraint sent	temperature 1.0, 1,024 max output tokens, on a non-thinking variant that takes no reasoning parameters	$0.09 / $0.55	the open-weights row in the census work (1 manifest)
`author-offplan`	`openai/gpt-5.5`	OpenRouter, `["openai"]`, fallbacks off	reasoning effort none unless a spec says otherwise	$5.00 / $30.00	authoring, verification and critic support; no run manifest records a call under this role name, since work of that kind that became paper-bound ran under `auditor`

separate them; Section 8 states the confound and Appendix G carries the results it qualifies. In the pipeline transfer run only the check stage moved to this family: extraction and audit were read from cache and are identical, character for character, to the primary-family run, and the second look stayed on the auditor model, so fact lists and severity grades are the same objects across both checkers. Checker reasoning effort is not matched tier for tier across families (in the primary family the two-stage pipeline ran at none and the three-stage at medium; the second family ran both at minimal), so deliberation is confounded with family in both directions there as well.

Models used outside the lock. Two, both outside the judging path: the census clustering's embedding model (named with its projection and clustering parameters in the companion census paper's method appendix (Fox et al., 2026)) and the text-to-speech renderer for the synthesised consultation audio, `gpt-4o-mini-tts`, named in the capture script rather than in the lock. Both sit outside the study's credit ledger: they were paid by card rather than credits and are not itemised in the study's artefacts. One further category needs stating rather than pinning: the three scribe products under audit are commercial systems whose internal models their vendors do not publish, so there is no pin to record for them and none is claimed.

D.2 Seeds

Replicates carry run seeds 11, 22 and 33. Within a run, each item takes `base_seed = run_seed x 100,000 + item_index x 10`, and sample *i* of an ensemble takes `base_seed + i`, so an eight-sample judgement occupies `base_seed+0` to `base_seed+7` and no two judgements can collide. Seeds that fix a selection rather than a sample are recorded the same way: the held-out confirmation subset was drawn at seed 41, the audit-stage severity validation pack at seed 20260817, and the clustering's dimensionality reduction at seed 42 with a stability sweep over 42, 43 and 44.

Seeds are accepted on the primary judge's route and on the second family's, and they do not make output reproducible at the token level. On the routing probe of 2026-07-30 the same prompt at temperature 0 and at temperature 1 returned different text on repeat calls while the verdict was stable in 4 of 4 probes. This is why the released raw completions, rather than the seeds, are what a replicator should check against: a rerun should reproduce our verdicts and our rates, not our strings.

D.3 The per-note prices, and when they were measured

The price ladder of Table K.6 is quoted per note, on the pinned routes of D.1, and every figure in it comes from the provider's own usage accounting for the runs named below rather than from a price list. These are what these models cost on those dates, and they move with any repricing.

price as quoted	what it buys	how it is composed	measured
$0.005	the engineered single-call judge, whose own critical-omission list is the cheapest flag in Table K.6 that names a fact	one call per note at the judge settings of Section 4	reference-judge run receipts, 12–13 August 2026
about $0.03 to $0.07	the RAGAS-style recipe with a calibrated threshold	$0.0305 per note measured, at 1.98 calls per note with the transcript fact extraction cached per consultation	cost read of 2026-08-14, over records bought 10–14 August 2026
$0.036	the eight-sample monolithic judge the two-stage pipeline is compared against	eight independent calls summing to $0.0360, against $0.0045 for a single call; the provider's prompt cache landed on all eight, so effective input cost is about 3.1 times a single call rather than eight	routing probe of 2026-07-30, at prices pinned 2026-07-29

price as quoted	what it buys	how it is composed	measured
$0.046	route two, the evolved prompt	one call per note at high reasoning effort under a 12,000-token cap	reasoning-budget run receipts, 18 and 24 August 2026
$0.094	the two-stage pipeline at one note per consultation	extraction $0.0868 per consultation plus $0.0077 per note for the single keyed check call, which is $0.0945	confirmation run receipts, 13–14 August 2026
$0.45	the three-stage pipeline at one note per consultation	extraction $0.0868 plus audit $0.3194 per consultation, plus $0.0440 per note for the check and the second look, which is $0.4502; the audit stage is 71% of that total	confirmation run receipts, 13–14 August 2026
$0.028 and $0.140	the same two tiers amortised over the 4.21 notes per consultation this corpus holds	$0.0283 and $0.1404 fully loaded	confirmation run receipts, 13–14 August 2026

Two conventions travel with those numbers. The one-note-per-consultation framing and the amortised framing must both be labelled wherever either is quoted: the extraction and audit stages run once per consultation and are cached, so the two framings differ by more than a factor of three on both tiers and describe different deployments, a practitioner evaluating one note per consultation paying the first and a service evaluating several notes against one transcript paying the second. And within the three-stage tier the second look accounts for $0.014 of the $0.044 marginal per-note cost, 32% of that tier's per-note spend, for no measurable change in detection in either direction (Appendix C).

The $0.03 to $0.07 range for the RAGAS-style recipe spans its two measured prices: $0.0305 per note with the per-consultation extraction cached across a consultation's notes, and $0.069 measured with the keyed contract and full token budgets (from the baselines runs' own receipts, released beside their manifests in the data repository).

E CLINICIAN VALIDATION IN FULL: BOTH SITTINGS

Every other measurement in this study is a model judging a model. Two sittings put a doctor under that stack. The first, on 14 August 2026, covered 70 items across six stages of the pipeline and the benchmark - 64 scored items across the six stages, plus three realism ratings and three free-text answers - and is the source of the adjudication reported in Section 7. The second, on 17 August 2026, covered 24 items and validates the one thing the first did not reach: the severity grades the deployable decision rule actually fires on. This appendix carries the blinding record, the per-stage rates, the stages that are measurements rather than endorsements, and the second sitting; it also reports the standing conflict: the rater is a physician author of this study who designed the instruments they are grading and knew the study's claims before they started.

Two rules govern how the two sittings are read together. First, they measure different instruments and their numbers are never pooled. The first sitting's severity stage validates the **benchmark's** severity grades, which are the axis every severity-conditioned result in Sections 5 to 7 is conditioned on; the second validates the **audit stage's own** grades of its own extracted facts, which is a separate run of the same written rubric and is what Section 7's decision rule fires on (that rule: flag the note if any fact graded critical is verdicted absent). Second, blinding differs by stage and the label travels with every rate. Where an item has to state the machine verdict in order to be answerable at all, the answer is an **endorsement** - a clinician checked and did not object - and where the item can be answered without the verdict on screen, the answer is a blinded **measurement**. Stages A, B, C and G of the first sitting are endorsements; D, E and F are measurements, and so is the second sitting's blind pass.

E.1 THE 70-ITEM SITTING, 14 AUGUST 2026

E.1.1 BLINDING, STAGE BY STAGE

stage	what it audits	how it was blinded	reads as
A	the deployed extractor's inference-time fact list	not blindable: the item prints the fact and its quote, which is the verdict under audit	endorsement
B	the reference-note repair loop	not blindable: the item prints the flag and the repair	endorsement
C	the omission-pair verification panel	not blindable: the item prints the rejection and its reasons	endorsement
D	the taxonomy panel's cuts	the audit classifier's four buckets were interleaved and never named on screen	measurement
E	the benchmark's severity rubric	no machine grade printed, and no indication of which items the two graders split on	measurement
F	the per-fact pipeline judge against the monolithic judge	neither which judge said what, nor whether the note shown was an original or an edited one	measurement
G	scenario realism and fact-sheet fidelity	a rating, not an adjudication; the rater designed the material	endorsement

E.1.2 THE PER-STAGE RATES

stage	stage under judgement	n	what agreement means	rate	Wilson 95%
A	the deployed extractor's inference-time fact list	10	the cited quote supports the extracted fact	10/10 = 100%	72.2–100%
B	the reference-note repair loop	8	the flag was a real defect and the repair fixed it	8/8 = 100%	67.6–100%
C	the omission-pair verification panel	6	the panel was right to reject the pair	6/6 = 100%	61.0–100%
D	the taxonomy panel's cuts	10	*false-kill rate*: a cut finding was a real omission	0/10 = 0%	0–27.8%
E	the benchmark's two-grader severity rubric	20	exact grade match	14/20 = 70%	48.1–85.5%

stage	stage under judgement	n	what agreement means	rate	Wilson 95%
F	the per-fact pipeline judge against the monolithic judge	10	the **pipeline's** per-fact verdict matches	10/10 = 100%	72.2–100%
G	scenario realism, extraction fidelity	3	a 1–5 rating, not an agreement	5, 5, 5	-

E.1.3 E: THE BENCHMARK'S SEVERITY GRADES

Twenty facts drawn from the benchmark's severity strata, graded blind against the written rubric of Section 3 (printed verbatim in Appendix A). The key is the consensus grade of the benchmark's two graders, which are two models from different families running the same rubric, with a split resolved to the lower grade.

- **Exact agreement 14/20 = 70.0% [48.1, 85.5].**
- **Weighted agreement, an adjacent grade counting half: 85.0%.**
- **Linear-weighted kappa 0.63**, where kappa is chance-corrected agreement and 0 is what two raters would reach by guessing (observed agreement 0.850, expected 0.595; percentile bootstrap 95% 0.32–0.86 over 10,000 seeded resamples, indicative at n=20). Unweighted Cohen's kappa 0.55.
- **No disagreement is more than one grade.** Six adjacent, zero two-grade.

Confusion matrix, rows the machine consensus, columns the clinician:

machine \ clinician	critical	supporting	peripheral
critical (7)	3	4	0
supporting (7)	1	6	0
peripheral (6)	0	1	5

The disagreement lives in the machine's critical row. Four of the six disagreements are the clinician moving a machine `critical` down to `supporting` (E1, E7, E9, E10); of the seven facts the two graders called critical they kept three. The other two run upward, one `supporting` to `critical` (E19) and one `peripheral` to `supporting` (E8). Their overall mix is 4 critical / 11 supporting / 5 peripheral against the machine's 7 / 7 / 6 - the same middle-heavy shape, with them the more conservative grader at the top. That is the opposite of the direction the rubric's resolve-downward rule was written to guard against. If the critical stratum is over-inclusive by anything like the four-in-seven this sample shows, every severity-conditioned result is conditioned on a slightly generous critical class. At seven machine-critical items that is something to check, not a correction to apply.

The items the two graders split on are the ones they match best, which is the reverse of the natural worry. On the 8 items where the graders disagreed and the tie was resolved downward, **7/8 = 87.5% [52.9, 97.8]**; on the 12 they agreed on, **7/12 = 58.3% [32.0, 80.7]**. Fisher's exact test gives two-sided p = 0.32, so this is not a significant difference and may well be noise, but it is worth recording because a reader will assume the split cases are the shaky ones. Here they are not.

Their three borderline comments all sit on items they **agreed** with, and two of the three say they would have accepted the adjacent grade. They are the qualitative version of the gap between 70% exact and 85% weighted:

- **E4** (machine critical, they said critical): *"should document those symtpoms, don't need each one exactly written literally, but it's quite critical to document that they were well overall and that that general vibe of things was asked about"*

- **E12** (machine peripheral, they said peripheral): *"could really be supporting or peripheral"*
- **E17** (machine critical, they said critical): *"i might accept supporting though too"*

The twenty facts, their machine-consensus keys, the clinician's answers and the exact/adjacent scoring are in the released sitting artefact, item by item. Some items grade a handling instruction rather than a bare fact, because for those the material at risk is defined by what a correct note would have recorded; the grading task is the same either way.

E.1.4 F: THE JUDGE DISAGREEMENTS, ADJUDICATED

Ten notes where the per-fact pipeline judge's verdict and the monolithic judge's note-level flag reach opposite conclusions, drawn one per consultation from the 35 such notes in the confirmation run (21 the pipeline flags and the monolithic judge does not, 13 the reverse, and 1 clean note the per-fact rule fires on). The monolithic judge here is the strongest of the eight designs, the completeness-scored eight-sample judge (`FC-score-k8` in the released data), read across three replicates. The clinician was told neither which judge said what nor whether the note in front of them had been edited, and was told explicitly that `?` was a real answer, because how much of this disagreement is genuine ambiguity is itself a finding.

They used ? zero times, and on all ten their verdict matches the pipeline's per-fact verdict and contradicts the monolithic judge's. Pipeline **10/10 = 100% [72.2, 100]**, monolithic judge **0/10 = 0% [0, 27.8]**, exact McNemar test on 10 discordant pairs **p = 0.002**.

The full adjudication table - each item's case type, the note's pair class and benchmark severity, the fact at issue, and all three verdicts (pipeline, monolithic across its three replicates, clinician) - is in the released sitting artefact (`validation/sitting_results.json`), item by item.

Two severity instruments are in play in that record and they must not be conflated. The benchmark severity above is the two-grader benchmark grade, which is the axis section E validates. The `critical` grade Section 7's decision rule actually fires on is the **pipeline audit stage's own** grade of its own extracted fact, a different run of the same written rubric, which is why F8 sits here as a `supporting` pair whose extracted fact the pipeline graded critical. Stage E does not validate the grades the decision rule fires on; it validates the axis the results are conditioned on. Nothing in this sitting puts a clinician under the audit stage's severity grading, which is what the second sitting in E.2 exists to do.

The six cases where the pipeline flags and the monolithic judge is silent are like-for-like, and they are the study's claim adjudicated on real cases. In all six the fact the pipeline named absent is the fact construction actually removed, the strongest monolithic design did not flag the note at all across three replicates, and the clinician confirms the fact is not in the note. A missing clindamycin allergy, a missing distal pulse check, a missing positive Murphy's sign: the monolithic judge read the whole note and scored it fine.

Two of the three reverse cases are not like-for-like and should not be quoted as clean losses for the monolithic judge. F4's note is a `change` pair and F5's carries a partial-strong omission of a fact the item never shows, so on those two the note-level flag can be earned by a defect the clinician was not asked about, while the pipeline is right about the specific fact. F2 is the clean one: an unedited note flagged in one replicate of three, with the clinician confirming the fact is present. Restricting to the **8 items where the monolithic judge's position is attributable to the fact in front of them**: pipeline 8/8 = 100% [67.6, 100], monolithic judge 0/8 = 0% [0, 32.4], p = 0.008. Both the 10–0 headline and the conservative 8–0 reading hold.

And the false alarm is not a false alarm. F6 is the confirmation run's single false positive under Section 7's per-fact rule, recomputed here from the three-stage pipeline's confirmation records (`confirm-B3.jsonl`) as 27 of 131 detection and 1 of 47 on clean notes, exactly the figures Section 7 reports, on the clean reference note for `aci|D2N199`. The flagged fact is "patient is aged 74". The transcript says it. The repaired reference note never states an age at all. The clinician's verdict is that the fact is not captured - a real omission of the note. But the rule's trigger is a fact graded critical, and this fact graded supporting in both of the clinician's passes (the second sitting's reconciliation, below), so under the rule's own trigger the flag stands as a false alarm, exactly as the body counts it. This is one case rather than a rate. It also says something section B's 8 of 8 does

not: the repaired reference notes are complete against their fact sheets, not against everything the transcript contains, and a judge that enumerates from the transcript will find the difference.

E.1.5 WHAT THIS VALIDATION CANNOT DO

1. **One rater, one sitting, no re-test.** There is no within-rater reliability figure and no way to separate their standard from the instrument's. Stage E's kappa is a two-rater figure in which one of the raters wrote the rubric.
2. **They are an author.** Stages A, B, C and G are endorsements of positions the items had to state to be answerable. D, E and F are the sections where blinding was structural and are the only ones that should be described as independent measurements.
3. **Every stage holds 6 to 20 items.** Four stages read 100% with lower bounds of 61 to 72%: that rules out gross failure and nothing finer. The intervals are the result; the point estimates are not.
4. **The sampling is seeded and stratified for coverage, not adversarial.** Stage A drew one fact per consultation with no filtering for interestingness and no fact the audit stage had rejected. Stage F drew one note per consultation stratified by stratum and by how much of the fact survived, not by which judge was likely to be right.
5. **The pack itself cost two answers.** Eight of the ten section D items printed a window on the note rather than the whole note, and both abstentions came from that. A rebuilt D section printing the full note would be the way to test the `present_elsewhere` finding properly (the audit classifier's bucket for cuts where the fact appears elsewhere in the note - a quarter of the census's 40 sampled omission refusals).
6. **The `present_elsewhere` bucket is unvalidated.** One answer and two abstentions across its three items, so nothing in this sitting supports or undermines the taxonomy's 25% share for it, and it must not be cited either way.
7. **Stage E does not reach the grades the decision rule fires on.** It validates the benchmark's severity axis, not the pipeline audit stage's own grades of its own extracted facts. That is the gap E.2 closes.
8. **Nothing here validates the omission-detection numbers themselves.** It validates the material they are measured on (A, B, C), the panel that produced the taxonomy rates (D), the severity axis they are conditioned on (E) and the judge design comparison (F). Detection and false-alarm rates remain what Sections 5 to 7 measured.

E.2 THE 24-ITEM AUDIT-STAGE SEVERITY SITTING, 17 AUGUST 2026

E.2.1 WHY A SECOND SITTING

Section 7's decision rule fires on the audit stage's own severity grade of its own extracted fact. The first sitting never scored those grades: it validated the benchmark's grades, which are a separate run of the same written rubric over different material. A second blinded pass covers the audit stage's grades directly. The two sittings are separate instruments and their numbers are never pooled.

E.2.2 THE PACK

Twenty-four audit-stage-graded facts from the 47-consultation confirmation run, one item per consultation, at deterministic seed 20260817 (pack `01db4358fb60`, builder `build_severity_pack.py`). The draw is deliberately weighted towards the class the rule depends on: **10 rule-fired facts** - graded critical by the audit stage *and* verdicted absent, drawn from the 31 such facts in the run, and including the `aci|D2N199` "patient is aged 74" clean-note false alarm - plus 5 random machine-critical, 5 random machine-supporting and 4 random machine-peripheral facts. Each item printed the full transcript. The machine grade was structurally blinded, and a byte-identity check over the stripped item cards verified that nothing on screen varied with the grade. Sat 2026-08-17, 16:06 to 16:14 UTC, 7 minutes 49 seconds, all 24 answered, zero abstentions, zero comments.

E.2.3 THE BLIND PASS

stratum	n	exact	weighted (adjacent = half)
rule-fired (all machine-critical)	10	**1/10 = 10% [1.8, 40.4]**	50%
random machine-critical	5	2/5 = 40%	70%
random machine-supporting	5	3/5 = 60%	80%
random machine-peripheral	4	3/4 = 75%	88%
machine-critical pooled	15	**3/15 = 20% [7.0, 45.2]**	57%
all 24 (stratified, critical-heavy)	24	9/24 = 38% [21.2, 57.3]	67%

Confusion matrix, rows the audit stage's grade, columns the clinician's blind grade:

audit stage clinician	critical	supporting	peripheral
critical (15)	3	11	1
supporting (5)	0	3	2
peripheral (4)	0	1	3

Their mix is 3 critical / 15 supporting / 6 peripheral against the audit stage's 15 / 5 / 4. Linear-weighted kappa is 0.28 and is **not comparable to the first sitting's 0.63**: this sample is deliberately critical-heavy, which biases kappa, so the per-stratum rows are the result and the kappa is not. There is one two-grade disagreement, S02 ("he had been coughing a lot for the last two days", machine critical against clinician peripheral), where the first sitting had none. The two near-duplicate distal pulse-examination facts, S14 and S16 from different consultations, both drew "supporting", which is internally consistent grading.

The random strata carry 4 and 5 items each and are quoted as counts for that reason; only the three rows with intervals above carry enough weight to read as rates.

E.2.4 THE UNBLINDED RECONCILIATION

The same day, the 15 blind disagreements were re-adjudicated unblinded, with the machine grade and the audit stage's recorded reasoning shown per item. **Five revised to critical** - S02 (the cough sitting beside calf findings, a pulmonary embolism picture), S03 (avoiding non-steroidal anti-inflammatories on renal risk), S07 (the blood-pressure safety-net), S11 (nil by mouth before possible surgery) and S23 (work of breathing in a febrile child) - all cases where the critical grade rests on consultation context that a fast blind pass had not weighed. **Seven stood** on the first reconciliation pass (S13 a bare age, S14 and S16 the normal pulse examinations, S20 family screening, and S08, S10, S06), and **three more stood on a follow-up confirmation** (S12, S17 and S22, all random machine-critical items the clinician kept at supporting).

On the pulse-examination pair they gave the reasoning verbatim:

> "if the pulses were absent i think it would be critical, but given they are present it's slightly more on supporting, but very borderline"

Reconciled: **rule-fired criticals confirmed 6/10 [31.3, 83.2]**, and **rule-fired facts graded clinically material - critical or supporting - 10/10**. Machine-critical pooled 8/15. Overall exact agreement 14/24 = 58%, weighted 79%. **Every residual disagreement is a single grade**, and the audit stage is the more severe grader in 9 of the 10 residuals; the single exception is S06 (a machine `peripheral` the clinician put at supporting).

All 24 items - stratum, the fact graded, the audit stage's grade, the blind answer, the reconciled answer and its status - are in the released validation artefact. Fact wording there is the audit stage's own, including the US trade names the ACI-Bench consultations use (Motrin is ibuprofen, Tylenol paracetamol).

E.2.5 WHAT THE TWO PASSES SAY

Blind, the audit stage's critical class looked substantially generous: 1 of 10 rule-fired criticals confirmed. **Reconciled against the audit stage's own reasoning, half that gap was consultation context the fast blind pass had not weighed**, and the final position is 6 of 10 rule-fired criticals confirmed, with the residual generosity being one grade on routine content - a bare age, normal examination findings, family screening. Both passes are reported, blind first.

The flags are clinically material. All 10 of 10 rule-fired facts are critical or supporting on reconciliation, so a flag names something a clinician would want in the note even where the "critical" label is one grade high.

Detection and false-alarm rates are unaffected. They are measured against constructed absence, not against severity truth. What changes is what a flag means: a material missing fact that the audit stage grades critical, with that grade sitting about one grade above this clinician's on routine content. Aligning the audit stage's critical threshold with clinician grades and re-scoring the already-purchased per-fact verdicts is a cheap lever, and the stored verdicts make it a re-scoring rather than a re-run.

The direction is the rater's as much as the instrument's. The same written rubric has since been graded blind by an independent clinician - not an author of this study and with no involvement in it - on a fresh sample of findings in the companion census (Fox et al., 2026): 9 of 12 grades exact and 3 one grade apart. Across the two census sittings 25 of 32 grades are exact and no disagreement anywhere exceeds a single grade, so the rubric reproduces to within a grade under two clinicians who graded disjoint material. What the pair does not establish is a direction, because the two clinicians lean opposite ways against it (Appendix A). The one-grade lean recorded above is therefore read as a property of this rater as much as of the audit stage, and no claim in this paper rests on the rubric leaning one way.

E.2.6 CAVEATS

One rater, an author, one 8-minute blind sitting plus a same-day unblinded reconciliation, and no re-test. The rows that carry the reading hold 10 and 15 items, so the intervals are wide and the intervals are the result. The reconciliation is unblinded by construction and is reported beside the blind pass, never instead of it. And the "patient is aged 74" fact (S13 here, F6 in the first sitting) drew supporting in both passes, so the first sitting's observation stands - the note really does lack a transcript fact - while the flag remains a false alarm under the rule's critical-only trigger, the supporting grade now confirmed twice.

Artefacts (paths in the released data repository's `validation/` directory unless noted). First sitting: `sitting_results.json` (the per-item join, per-stage rates with Wilson intervals, the severity confusion matrix and kappa, the adjudication table, the position answers verbatim), the raw answer export `sitting_answers.json`, and the deterministic builder `build_sitting_pack.py` in the code repository - the pack itself prints withheld note excerpts and is not released. Second sitting: `build_severity_pack.py` (code repository), `severity_answers.json`, `severity_pack_keys.json` and `severity_validation_results.json`.

F RELEASE INDEX

The reproducibility statement states what we release and the ethics statement what we withhold. This appendix names the release's two homes and its licences, sets out the withholding policy, and carries the compressed datasheet - composition, verification standards and known limitations. The field-level schemas, the scoring protocol and the usage guide live in the bundle's dataset card, which ships as the data repository's `README.md`, and nothing in the card claims more than those two statements do.

The benchmark is released in two parts. The data - everything the datasheet below describes - is published under CC BY 4.0 and carries a DataCite DOI minted at publication against the released revision. The harness is published under the MIT licence and archived to a preservation repository at its tagged release, so the code has a citable snapshot independent of the hosting account. The release's name and both repository addresses are withheld here because they identify the authors, and

because a link whose host logs visitors could expose reviewer identity; the addresses are published when the submission is de-anonymised. The file-by-file index of the release is withheld with them, since it is unusable without the addresses; everything a reader needs to judge what the benchmark is, and what is wrong with it, is in the datasheet below.

Anonymisation of the harness itself. The study was not run anonymously, so some of its file names and configuration keys carry product names. The packaging step rewrites them to the letters and then fails the build on any case-insensitive product-name match anywhere in the bundle. A released capture script therefore names its driving mode - a documented API, or audio played to the product in real time - and never the product; the letter-to-product mapping stays in the private repository and is not released in any form.

F.1 What we withhold

What we withhold is the note text the three products wrote. The reason is contractual rather than clinical: the corpus contains no patient data at all (the ethics statement), and it is the products' terms of service, which differ from one another, that decide whether their output can be republished. We treat all three the same way rather than release asymmetrically, partly because a release carrying one product's notes and not another's would make products identifiable by their absence.

Three things keep the withheld half honest. Released findings carry the transcript-side evidence for every claim and quote the note's own words only in the ten internal-contradiction findings that cannot be stated without them, and the packaging step fails the build on any released content that reproduces a run of withheld note text, on any schema field carrying a note body, and on any product name anywhere in the bundle - the checks themselves are described in the released `VERIFICATION.md`. A replicator with their own account can regenerate equivalent notes end to end for the API-driven product, once the identifiers the anonymisation removed are restored; for the other two the capture scripts are published as the record of how the capture ran rather than as a route anyone can re-run. And researchers who need more than the release carries can contact the authors.

F.2 The benchmark datasheet, compressed

The full datasheet is the bundle's dataset card; this section carries what a reader of the paper needs - composition, verification and limitations. Intended use is evaluating LLM judges and ambient scribe systems; it is explicitly not for clinical use of any kind or as training data for clinical systems, and nothing in it is a real patient encounter or a clinical record.

F.2.1 Composition and construction

Composition is Table A.1's job, in A.6, and construction is Appendix A's; this datasheet does not repeat them. In one sentence each: the corpus is 145 consultations over four strata (30 authored, 10 trap-blind with 9 kept, 57 PriMock57 with 53 kept, 48 ACI-Bench with 45 kept), of which 134 carry a verified clean note - 112 evaluation, 22 in a separate judge-optimisation split - and the released set is 500 pairs, 495 in the evaluation set - 293 omissions (150 complete, 86 fragment trace, 57 restatement trace) and 202 commissions - plus the 112 clean twin notes they are scored against. Construction runs transcript-first through nine steps - deterministic subsample (seed 20260728, the 67 previously-used ACI-Bench encounters excluded), blind extraction, three-critic audit, consolidation, reference-note repair, site mapping, single-error injection, the optimisation-split carve, and cross-family verification with failures excluded rather than patched - shown with its counts in Figure A.2. The repaired reference note ships inside each pair record as the clean note; the originals belong to the upstream corpora's own releases; the 27 blind re-extractions of authored transcripts are an instrument-comparison artefact from which no pair is built.

F.2.2 Verification standards

Nothing in the dataset is described as clean by construction. Each layer has an instrument, a measured rate on a stated denominator, and a published failure list.

layer	standard	measured
fact sheets	three critics, one revision cycle, drop if material issues remain	sheets dropped: PriMock57 4 of 57 (7.0%) after three disclosed instrument corrections, ACI-Bench 3 of 48 (6%), trap-blind 1 of 10, authored re-extraction 3 of 30
fact-sheet fidelity	blind extraction against the authored sheet on the same 30 transcripts	646 of 650 authored facts recovered, 99.4% [98.8, 99.9], per-consultation minimum 93.8%
reference notes	audited against transcript and core sheet, repaired, re-verified	53 of 53 PriMock57 and 45 of 45 ACI-Bench notes carry at least one material discrepancy with their own transcript, at means of 10.7 and 7.1 per note; 17 of the 98 repaired notes still fail an ultra-strict residual check and are logged
clean notes	four-section checklist audit by a different model family, key in hand	first-pass defect rate by stratum: authored 11 of 30 (36.7%), trap-blind 3 of 9 (33.3%), PriMock57 49 of 53 (92.5%), ACI-Bench 36 of 45 (80.0%); 88 notes repaired, 3 excluded
pairs, earlier build	deterministic single-edit check, then a semantic edit check at three seeds with a per-field majority	first-pass failure 182 of 402 (45.3%) [40.5, 50.2]; 78 of 411 pairs excluded
pairs, graded-omission build	cross-family panel of two model families, third call on any field they split on, per-field majority	verified 71 of 126 complete-removal attempts (56.3%) and 114 of 151 partial removals (75.5%); 92 pairs dropped, itemised with their failing checks
severity grades	written rubric, two independent cross-family graders, conservative tie-break	Cohen's kappa 0.662 [0.59, 0.73] over the 683 traps regraded under the rubric, against 0.177 [0.04, 0.31] over 171 matched traps graded without it, intervals non-overlapping; 92 of the 683 (13.5%) resolved to the lower grade

Two disclosures belong with these numbers. The audit rates are auditor-majority rates with no human adjudication in the loop, so a defect rate is an upper bound on a human-confirmed one; the empirical check on auditor strictness is that 15.3% of one-seed screen flags die at three-seed majority confirmation (57 of 373), rising to 29.6% of the 98 flags in the open-ended section. And audit reproducibility is 0.794 over the escalated checklist sections and 0.725 including the open-ended one, both conditional on the screen having flagged something at all (222 of 1,024 section-audits escalated), and therefore a lower bound over disputed sections rather than a whole-corpus figure.

F.2.3 Known biases, gaps and exclusions

Population and setting. English-language UK primary care and US ambulatory care only. Two of four strata are consultations we wrote; the largest public stratum is UK primary care with actor patients. No secondary care, no specialty outside general practice, and no note in the corpus carries an electronic-record structured field, so the benchmark cannot speak to structured-field omissions at all.

No real patients anywhere. Actors, simulation and fiction throughout. That is an ethical strength and an ecological limitation at once: acted and simulated consultations are cleaner and more purposeful than real ones.

Audio provenance is mixed. One stratum is genuine recordings, three are synthesised speech, and the recordings predate the products by several years. All capture ran at native speed on unmodified audio, after a speed-up test was rejected when silence-trimming produced a critical medication substitution. The audio itself is regenerable from the released transcripts with the harness's text-to-speech recipe and pins.

Selection asymmetry between strata. First-pass pair failure was 18.9% on the authored stratum against 52.9% on PriMock57 and 52.3% on ACI-Bench, because the authored stratum is preserved verbatim and its failing pairs are excluded rather than regenerated, while public-stratum notes were repaired and pairs regenerated up to three times. A surviving public-stratum pair has passed a harsher filter, so judge and scribe scores should not be compared across strata.

Two construction cohorts are pooled, and recorded per row. The 202 commission controls and 79 complete omissions carried over were built by model rewrite rather than by span deletion off the site map, and the 79 carry severity grades from an older single-arm grading; the 34 relabelled partial seeds vary in residual count from 1 to 9 because they predate the site map. Every pair records which cohort it came from, so a user can restrict to the internally uniform column - noting that all commission controls sit in the carried-over cohort, so the uniform column has no commission arm.

Residual count is a property of the map, not of the note. Construction left exactly one mapped residual site on all 151 partial pairs, but the verification panel's own reading found more than one surviving mention in 71 of them. Both counts are in each record, so an analysis can control for it; the released trace labels use the map's count.

Thin and unpopulated cells. The restatement-trace peripheral cell holds 2 pairs and should be treated as unpopulated, and the fragment-trace peripheral cell is thin at 8. A restatement trace needs a fact whose second mention is explicit or a full paraphrase, and peripheral facts are both the rarest in the corpus and the least redundant.

Excluded content, with counts.

what	count	why
consultations dropped at the fact-sheet gate	8 of 145	material issues remained after one revision cycle; Figure A.2's 11 of 145 counts sheets, adding the 3 authored consultations whose blind re-extraction dropped while their own authored sheets stand
clean notes excluded	3 of 137	never converged after three repair rounds
pairs excluded at the earlier build's answer-key audit	78 of 411	27 omit pairs failed three regenerations and the deletion fallback, 20 add and 5 change failed three regenerations, 17 authored pairs failed the semantic majority and are never rewritten, 9 were orphaned by the excluded notes
graded-omission pairs dropped for failing verification	92 of 277	itemised with their failing checks in the released dataset
encounters tagged as overlapping an external benchmark	14 of 48	retained in the benchmark, excluded from any analysis that also uses that benchmark

Everything is model-produced except the transcripts. Fact sheets, clean notes, injections, repairs, site maps and audits are all model outputs, checked by models from a different family. Human clinical review is by an author-clinician and is disclosed as such; its scope and its limits are in Appendix E.

Anonymisation and safety, licences, versioning and citation are F.1's and the head of this appendix's job, restated in full in the dataset card: the three systems are Scribe A, B and C everywhere, the released set carries `dataset_version: factorial-v1` (built 2026-08-12, superseding the

earlier 281-pair freeze whose digests identify the carried-over cohorts' source), the holdout from the prompt optimiser is asserted structurally at build time, and a number computed over the benchmark should be quoted with the dataset version, because a later revision mints its own DOI and the two are not interchangeable.

F.3 PER-CORPUS ATTRIBUTION

Both public corpora are published under CC BY 4.0, and both are redistributed here in derived form. The transcripts are normalised, and for ACI-Bench subsampled. The clean notes for those two strata are the corpora's own reference notes audited against their transcripts and repaired, so every pair built on them is an adaptation of upstream material, not only of ours. The fact sheets derived from those strata are adaptations too.

CC BY 4.0 requires four things of a derived release: name the creator, carry the licence with a link to it, state that changes were made and what they were, and impose no further restriction on downstream users. The release carries one fixed attribution string per corpus - naming the creator, the licence, the source record, and the changes made (transcripts normalised; for ACI-Bench a 48-encounter subsample drawn at seed 20260728; the clinician reference notes audited against their own transcripts and repaired; single-error variants generated from the repaired notes) - reproduced verbatim in the bundle's `ATTRIBUTION.md` beside the licence file, in the dataset card, and in the `_attribution` header of every derived JSON artefact. For ACI-Bench the CC BY 4.0 grant is carried on the corpus's figshare record, which the string cites; for PriMock57 it is the repository's own licence file. Everything else in the bundle is our own work released under CC BY 4.0, so a downstream user carries one licence across the whole bundle and two attribution lines.

Both upstream licences were verified against their published sources at packaging time, and CC BY 4.0 expressly permits redistribution of adapted material with attribution, which is what this bundle is.

G THE SECOND JUDGE FAMILY, IN FULL

This appendix carries every number behind the cross-family sentences in Section 5 and the second limitation in Section 8: what was run on a second judge model family, on which sets of pairs, and what came back.

Its scope should be read before its numbers. This is one additional judge model family, and on it we ran two of the eight judge designs as monolithic judges plus one replicate of each pipeline tier with only the per-fact checker swapped. It is not a second ablation and it does not license family-versus-family comparisons. The completeness-scored single-sample design carries three replicates here; the faithfulness-only yes/no design carries one, and so does each pipeline tier. The replicates establish that each family's own figure is stable from run to run, not that the difference between the families is real: no test of that difference was run, and none of the deltas below is a tested effect. The eight-sample completeness-scored design was never run on this family, so two designs of the eight are the whole monolithic picture.

G.1 WHAT WAS RUN

The second-family judge is `google/gemini-3.1-pro-preview`, served through OpenRouter on the `google-vertex` route, at temperature 1.0, `max_tokens` 1024, and `reasoning_effort minimal`. The endpoint refuses to run with reasoning switched off, which is the first of this appendix's two confounds (Section G.4). The primary family is the pinned GPT-family judge of Section 4, at the settings it ran everywhere else. Appendix D carries the model pins, and names the release homes of the prompt hashes and run manifests for both runs here.

The monolithic replication used the judge prompts byte for byte as the eight designs ran them, on the same set of pairs: 495 evaluation pairs and 112 verified-clean twins, giving 293 omission records, 202 commission records (additions and alterations) and 112 clean twins per design per replicate. Replicate seeds are the same three the eight designs used (11, 22, 33). Every primary-family row printed beside a second-family row below is recomputed from the original judgement store by the

same scoring functions and checked against the published figures before anything here was written, agreeing within 0.0015 on all four paired values.

G.2 The asymmetry and the restatement-trace collapse both reproduce

Paired discrimination is the share of pairs in which the judge scores the flawed note strictly below its own verified-clean twin, plus half of any ties, so 0.500 is a coin flip. Both error columns are the same measure on the same notes; only the class of error differs.

design	judge family	replicates	paired commissions	paired omissions	commissions minus omissions
faithfulness only, yes/no, asked once (F-bin-k1)	primary	3	0.875	0.500	+0.375
faithfulness only, yes/no, asked once (F-bin-k1)	**second**	1	**0.951**	**0.551**	**+0.399**
faithfulness and completeness, 0-to-10 score, asked once (FC-score-k1)	primary	3	0.895	0.585	+0.310
faithfulness and completeness, 0-to-10 score, asked once (FC-score-k1)	**second**	3	**0.961**	**0.683**	**+0.278**

Multi-replicate rows are the mean of the per-replicate values. The direction, the rough size and the ordering of the two designs all reproduce: on this family, content that was added or altered is caught at 0.951 and 0.961 while missing content sits at 0.551 and 0.683, a gap of 0.40 and 0.28 on a measure whose chance level is 0.500.

Every paired and detection figure in the table below sits above its primary-family counterpart, at a lower false-alarm rate, and the per-replicate rows show that the completeness-scored design's own figure is stable rather than a single lucky draw. The exception among the monolithic rows is the restatement-trace case, G.2's residual table; G.3's pipeline transfer has several figures that move the other way. Each row is a complete independent pass over the full set of pairs; nothing is pooled inside a row. Detection is the flag rate on the 293 omission-errored notes under the score threshold fixed before the run (a score of 7 or below), and the false-alarm rate is the same design's flag rate on the 112 clean twins.

replicate	seed	paired commissions	paired omissions	omission detection	false alarms (clean twins)	detection-vs-false-alarm z
r1	11	0.960	0.700	20.1% (59/293)	1.8% (2/112)	4.62
r2	22	0.965	0.675	19.1% (56/293)	0.9% (1/112)	4.72
r3	33	0.958	0.674	20.5% (60/293)	3.6% (4/112)	4.17
mean / pooled		**0.961**	**0.683**	**19.9% (175/879)**	**2.1% (7/336)**	**7.79**

Across the three replicates: paired omissions 0.674 to 0.700, paired commissions 0.958 to 0.965, detection 19.1 to 20.5% against a false-alarm rate of 0.9 to 3.6%, and a detection-versus-false-alarm

separation that is significant in every one, z 4.17 to 4.72. On this family the completeness-scored single-sample judge clears its own noise on single notes in all three replicates, which no design in the primary family did.

Conditioning on how much of the fact survives reproduces as well, including the part that matters clinically. A restatement trace means the fact's main statement is gone while an explicit or closely paraphrased mention survives elsewhere in the note.

replicate	complete removal	fragment trace	restatement trace
r1	0.792	0.657	0.518
r2	0.750	0.651	0.509
r3	0.770	0.622	0.500
mean	**0.771**	**0.643**	**0.509**

The restatement-trace case sits at chance in every replicate, 0.500 to 0.518, exactly as it does on the primary family, where the best of the eight designs reads 0.526. The severity breakdown runs in the same order on this family: critical 0.754, supporting 0.614, peripheral 0.590 for the completeness-scored design over three replicates, against the primary family's 0.683, 0.586 and 0.568 on its best design. The faithfulness-only design, on its single replicate, reads 0.567, 0.558 and 0.500 by residual and 0.570, 0.539 and 0.513 by severity.

G.2.1 Absolute detection, and how far the pooled separation can be trusted

Detection means nothing except beside the same design's false-alarm rate at the same threshold, so both are printed together, with the flag rule that produced them.

design	judge family	flag rule	omission detection	commission detection	false alarms (clean)	detection-vs-false-alarm z
F-bin-k1	primary, 3 reps pooled	FAIL verdict	9.6% (84/879)	84.0% (509/606)	8.9% (30/336)	0.34
F-bin-k1	**second, 1 rep**	FAIL verdict	**13.0% (38/293)**	92.6% (187/202)	**2.7% (3/112)**	**3.07**
FC-score-k1	primary, 3 reps pooled	score <= 7	4.7% (41/879)	58.7% (356/606)	5.7% (19/336)	-0.71
FC-score-k1	**second, 3 reps pooled**	score <= 7	**19.9% (175/879)**	90.9% (551/606)	**2.1% (7/336)**	**7.79**

The pooled z values in that table, on both families, count three passes over the same notes as three independent observations, which inflates them. The per-replicate z values above (4.17 to 4.72) are the honest figures for the second family, and the same caveat applies wherever a pooled row is quoted for the primary family. Nothing in the paper rests on a pooled z.

One further operating point needs its mechanism stated with it. Swept to the most sensitive threshold whose false-alarm rate stays within 10%, the second family's completeness-scored design reads 44.9% detection (394/878) at 9.6% false alarms (32/335), against 4.7% (41/879) at 5.7% (19/336) for the same design on the primary family. That 44.9% is bought at a threshold of "anything short of a perfect 10 is a flag", and it works only because this judge scores clean notes at exactly 10. It is an operating point on one corpus with no evidence that it survives off-distribution, and it is why the paper's cross-family claim is about the gap between commission and omission detection rather than about absolute levels. The swept denominators are 878 and 335 rather than 879 and 336 because the two records that failed to parse (below) drop out of the swept curve, one omission record and one clean twin.

G.2.2 Parse failures

Two records failed to parse in 2,428, both on the same consultation, `aci|D2N086`: none on the faithfulness-only design, two on the completeness-scored one. Twenty-two samples hit the single

retry fixed before the run and twenty of those recovered. Unparseable output is recorded as a parse failure and never halts a run.

In replicate 1 the failure was the clean twin, so every pair of that consultation loses its reference and drops out of the paired measure for that replicate. In replicate 2 it was an errored note, so that one pair drops. Absolute rates are unaffected, because a parse failure counts as not-flagged against the full denominator. Paired denominators after those drops: the faithfulness-only design 293 omission and 202 commission records; the completeness-scored design 290 and 200 in replicate 1, 292 and 202 in replicate 2, and 293 and 202 in replicate 3, against 293 and 202 with nothing dropped.

G.3 The pipeline transfer: only the check stage changed

The second run swapped the per-fact checker of Section 7's pipeline to the same second-family judge and changed nothing else. Extraction and the critic audit were read from the cache of the original run for all 47 consultations, so the fact lists and the severity grades are identical character for character across the two checkers; the quote-verified second look also stayed on the auditor model. Extraction runs on a third model family and the audit on the primary family, so in both configurations the enumerated facts are not the checking judge's own. The check stage ran at `reasoning_effort minimal`, temperature 1.0, one replicate, seed 11.

The set of pairs is Section 7's confirmation set exactly: 151 held-out pairs over 47 consultations (10 additions, 10 alterations, 62 complete omissions, 69 partial omissions) plus 47 clean twins, 198 notes per tier. The primary-family columns are recomputed from the original confirmation store by the same functions, and are replicate one throughout, so that both checkers are read at equal treatment; Section 7.1's 0.786 and 0.762 are means over three primary-family runs. There were zero parse failures on either tier.

tier	checker	paired omissions	paired commissions	swept detection at false alarms <= 10%
two-stage pipeline (B2)	primary	0.801	0.650	16.0% (21/131) at 8.5% (4/47)
two-stage pipeline (B2)	**second**	**0.798**	**0.575**	**19.1% (25/131) at 8.5% (4/47)**
three-stage pipeline with the severity audit (B3)	primary	0.752	0.500	18.3% (24/131) at 8.5% (4/47)
three-stage pipeline with the severity audit (B3)	**second**	**0.824**	**0.700**	**14.5% (19/131) at 4.3% (2/47)**

Conditioning by how much of the fact survives, paired, omissions only:

tier	checker	complete removal	fragment trace	restatement trace
two-stage (B2)	primary	0.935	0.792	0.561
two-stage (B2)	**second**	0.919	0.778	**0.591**
three-stage (B3)	primary	0.863	0.819	0.470
three-stage (B3)	**second**	0.911	0.847	**0.636**

The decision rule of Section 7 flags a note if any fact the audit graded critical is verdicted absent. Because the severity grades are the cached ones, the rule's trigger class is identical for the two checkers by construction, and what changes is only which facts each checker verdicts absent.

metric, on 151 held-out pairs and 47 clean twins	primary checker	second-family checker
detection, all omissions	20.6% (27/131)	**22.1% (29/131)**
false alarms, clean twins	2.1% (1/47)	**2.1% (1/47)**

metric, on 151 held-out pairs and 47 clean twins	primary checker	second-family checker
detection-vs-false-alarm z	2.99	**3.14**
detection on critical-severity pairs	32.8% (19/58)	**36.2% (21/58)**
complete removals	33.9% (21/62)	35.5% (22/62)
fragment trace	16.7% (6/36)	19.4% (7/36)
restatement trace	0.0% (0/33)	**0.0% (0/33)**
commissions flagged	15.0% (3/20)	15.0% (3/20)

The comparator that makes this a like-for-like test is the same family's own monolithic judge, restricted to exactly these 151 pairs and 47 clean twins from the records of Section G.2. Those records were already bought, so this row cost nothing and is measured on identical notes.

configuration, all judged by the second family	paired omissions	paired commissions	detection	false alarms
monolithic completeness-scored, asked once, 3 reps (score <= 7)	0.668	1.000	16.5% (65/393)	2.1% (3/141)
two-stage pipeline (B2), swept at false alarms <= 10%	0.798	0.575	19.1% (25/131)	8.5% (4/47)
three-stage pipeline (B3), swept at false alarms <= 10%	0.824	0.700	14.5% (19/131)	4.3% (2/47)
three-stage pipeline with the per-fact critical rule	(as B3)		**22.1% (29/131)**	**2.1% (1/47)**

The monolithic comparator's three replicates on this subset read 0.688, 0.658 and 0.660 paired on omissions, 1.000 on commissions in each, and 16.0% (21/131), 16.8% (22/131) and 16.8% (22/131) detection at 2.1% (1/47), 0.0% (0/47) and 4.3% (2/47) false alarms. The pipeline rows are one replicate each, so the monolithic mean is the like-for-like figure.

Five readings follow, and the first is the one Section 7 rests on. **Restructuring the task keeps its lead on the stronger judge.** On identical notes, the pipeline reads 0.798 and 0.824 paired against 0.668 for that family's own monolithic judge, worth +0.129 and +0.156.

The decision rule is the closest thing in this study to an operating point that is not judge-specific. It is a fixed rule over per-fact verdicts and was tuned on nothing, and across the change of checker family it moves from 20.6% (27/131) detection at one false alarm in 47 to 22.1% (29/131) at the same one false alarm in 47, which is two further true detections in 131 notes. On the 58 pairs carrying a critical-severity omission it goes from 32.8% (19/58) to 36.2% (21/58). At that matched 2.1% false-alarm rate this family's monolithic judge reads 16.5% (65/393), so the rule's edge here is +5.6 points of detection, alongside the fact that its flag names the missing fact and its severity where a monolithic flag is a number.

The ranking between the two tiers is judge-dependent even though the ranking of pipeline over monolithic is not. The three-stage tier overtakes the two-stage one on this checker (0.824 against 0.798), reversing the primary family's ordering, and the reason is visible in the verdict mix: this checker uses the trichotomy's middle label, which the primary checker at no reasoning effort did not. The same behaviour shows up downstream, where the quote-verified second look ran over 319 flagged absences on this checker against 299 on the primary family. Tier comparisons therefore have to travel with their checker.

The restatement-trace problem splits in two, and this part is new rather than a replication. On the paired measure the second-family three-stage pipeline discriminates restatement traces at 0.636,

well clear of the 0.47 to 0.56 band every method in Section 7 sits in; under the per-fact rule its detection there is 0 of 33, identical to the primary checker's 0 of 33. The information is present at the fact level and the decision layer cannot yet act on it, so the open problem of Section 7 is at least partly a decision-rule problem rather than purely a discrimination one.

The mirror image persists, and it is wider on this family. The monolithic judge scores 1.000 on the subset's 20 commission pairs while the pipeline tiers score 0.575 and 0.700 and the rule flags 3 of them: a judge that enumerates what should be present wins on absence and loses on invention here too. Twenty commission pairs make that size indicative rather than precise.

G.4 The two confounds, and what these runs cannot settle

The reasoning setting moves with the family. The second family's endpoint refuses to disable reasoning, so it ran at the minimal setting where the eight designs ran with none. Judge family and reasoning setting cannot be separated by these runs, and every second-family figure above carries that.

Checker effort is not matched tier for tier. In the primary family the two-stage pipeline's check ran at no reasoning effort and the three-stage pipeline's at medium, as Appendix C records; the second-family checker ran at minimal on both tiers. Family is therefore confounded with deliberation in both directions across the tier rows, and the tier reversal above is one place that matters.

Three further limits belong with the numbers. The pipeline rows are one replicate against a three-replicate monolithic comparator, and the faithfulness-only monolithic row is a single draw whose deltas are one run against a three-run mean. With 151 pairs the residual subgroups are small (62 complete removals, 36 fragment traces, 33 restatement traces) and their intervals are correspondingly wide; the commission side holds 20 pairs. And no significance test was run between the families, so the honest summary of this appendix is that the asymmetry, its conditioning, the restatement-trace collapse and the pipeline's lead over monolithic judging all reproduce on a second family, that single-note detection levels move with the family, and that a full cross-family comparison of the eight designs remains future work.

H The deployed faithfulness judge, verbatim

Sections 4 to 6 use a faithfulness judge deployed in our own production evaluation ("the deployed judge" in the floats) as a reference judge, in three wordings that differ only in how they treat omissions. This appendix prints the as-shipped prompt in full, so the wording experiment of Section 6 can be read against the exact text it edited, and records how the other two wordings differ from it. All three files ship in the release with SHA-256 hashes (`w2_prompts/PROMPTS.sha256`); the placeholders `{transcript}` and `{summary}` are filled per note at call time. The print below normalises typographic dashes, quotes and arrows to ASCII for the monospace font; the released files keep their original bytes, and the hashes refer to those.

H.1 What the three wordings are

The as-shipped prompt below carries an explicit exclusion: an "Omissions" bullet stating that information present in the transcript but absent from the note is not an error and must not be flagged, plus one mention of omission in a list of things not to flag. The **exclusion-deleted** wording (`v14_noexcl.txt`) removes exactly those two hunks and changes nothing else. The **affirmative** wording (`v14_incl.txt`) starts from the exclusion-deleted text and adds one bullet to the list of flaggable error types: a clinically significant omission - a finding, diagnosis, medication, allergy, safety-netting instruction, or red flag that is present in the transcript and material to the patient's care, but absent from the note. Section 6 reports what each change does: deleting the exclusion does not help and if anything hurts, and the whole effect belongs to the affirmative instruction.

H.2 The as-shipped prompt

```
### Task
You are a clinical safety reviewer. You are given a doctor-patient
```

```
conversation transcript and an AI-generated clinical note (a SOAP-style
summary) produced from that transcript. Decide whether the note
contains at least one factual error that a clinician reviewing this
note should be alerted to.

Output a single binary decision: YES (flag for review) or NO.

### The fundamental question

For each thing in the note that's worded differently from the
transcript, ask: "would a clinician form a meaningfully different
understanding of the patient from this note than the transcript
actually supports?"

If yes for any single item in the note -> flag YES.
If no for every item -> flag NO.

### What COUNTS as flag-worthy

Flag the note if it contains at least one of:

1. **Fabricated content** -- a fact, finding, value, name, dose, date,
   medication, or symptom in the note that is not in the transcript at
   all and cannot be reasonably inferred from context.

2. **Material misrepresentation** -- information from the transcript
   that the note got materially wrong: direction flipped (patient
   denied X, note asserts X), symptom substituted (patient said
   dizziness, note says nausea), attributed to the wrong subject
   (uncle's history attributed to father), or wrong value/quantity.

3. **Substantive over-reach** -- a clinical conclusion or specification
   in the note that goes beyond what the transcript supports in a way
   that changes clinical meaning. Examples: pre-supposing a diagnosis
   the doctor was actually evaluating; specifying a duration or
   quantity the patient did not give; adding a qualifier that narrows
   or shifts clinical interpretation.

### What does NOT count as flag-worthy

The following are NORMAL parts of clinical documentation. Do NOT flag
on these grounds, even though the wording differs from the transcript:

- **Paraphrasing into clinical language.** Patient says "my back hurts
  when I bend over" -> note: "back pain worse with flexion". Normal.

- **Standard medical abbreviations.** BP, HR, ROM, PMH, SOB, NKDA,
  WNL, etc. -- these are not errors when applied to transcript content.

- **Standard normal-exam phrasing.** If the doctor said "heart sounds
  clear" or "lungs clear", the note may expand to "no murmurs, gallops,
  or rubs" / "no wheezes, rales, or rhonchi" / "clear to auscultation
  bilaterally". This is conventional clinical write-up, not over-reach.

- **Standard imaging/lab summary statements.** Doctor describes
  specific normal findings on a study; note writes "unremarkable" or
  "within normal limits" -- conventional summary, not error.

- **Section headings and boilerplate.** "Chief Complaint:",
  "Assessment:", "Plan:", "Review of Systems:". The standard ROS
  template phrasing ("denies fevers, chills, weight changes...") when
  the patient was asked about general systems is also boilerplate.

- **Omissions.** Information present in the transcript that the note
  didn't include. Omissions are NOT errors. Flagging is for things
```

```
  that are wrong, not things that are missing.

- **Reasonable contextual inference.** Writing "today's visit" when
  the visit context makes the date obvious; writing "COVID vaccine"
  when the patient mentions getting "the vaccine" in a context where
  that's the relevant one; using the standard clinical name for an
  exam the doctor described informally.

- **Light synthesis.** Drawing a conclusion that follows directly from
  facts already stated in the transcript, where a clinician reading the
  note would arrive at the same conclusion.

Important calibration: most clinical notes contain MANY paraphrases,
abbreviations, summary statements, and template phrases. These are
normal. A flag-worthy note has at least one item that goes beyond
normal documentation -- not just "the wording differs from the
transcript". If you can't point to a specific item where a clinician
would say "wait, that's not what the transcript supports", the answer
is NO.

### Examples (synthetic -- illustrative)

Each example shows a transcript snippet and a corresponding note.
The verdict is the doc-level YES/NO decision and the reasoning.

--- Example 1 (verdict: NO) ---

Transcript:
[doctor] How have you been since last year, Mr. Park?
[patient] Pretty good. The blood pressure medication is working -- at
home it's been around 130 over 80 most days.
[doctor] Any chest pain, shortness of breath, or palpitations?
[patient] No, none of those.
[doctor] How's your sleep?
[patient] Sleep is fine.
[doctor] Let me take a listen. ... Heart sounds clear, lungs are
clear. You're up about two pounds from last visit -- that's within
normal variation.
[doctor] Let's continue the same medication and see you in a year.

Note:
Mr. Park presents for annual follow-up of hypertension. He reports
good control on his current regimen with home BP averaging 130/80.
He denies chest pain, dyspnea, or palpitations. Sleep is adequate.

Cardiovascular: No murmurs, gallops, or rubs.
Pulmonary: Lungs clear to auscultation bilaterally.
Weight: Up 2 lbs from prior visit, within normal variation.

Plan: Continue current antihypertensive regimen. Annual follow-up.

Verdict: NO
Reasoning: Every claim has a transcript anchor. "Dyspnea" is the
clinical term for "shortness of breath" -- paraphrase, not error. "No
murmurs, gallops, or rubs" is the standard write-up of "heart sounds
clear" -- light synthesis, not over-reach. BP and lbs are standard
abbreviations. No item where a clinician would form a different
understanding from the note than the transcript supports.

--- Example 2 (verdict: NO) ---

Transcript:
[doctor] What brings you in today?
[patient] My right knee has been bothering me for a couple of weeks,
mostly when I go up stairs.
```

```
[doctor] Any injury you remember?
[patient] No, it just started.
[doctor] Any swelling or redness?
[patient] No.
[doctor] Let me take a look. ... Tender on the inside of the knee, no
swelling, range of motion is full but with some discomfort.
[doctor] Could be a strain or early arthritis. Let's try ibuprofen
and physical therapy. If it's not better in a month, we'll image it.

Note:
HPI: Patient presents with two weeks of right knee pain, worse with
stair climbing. No history of trauma. Denies swelling or erythema.

Exam: Right knee tender to palpation medially. No effusion. Full ROM
with mild discomfort on flexion.

Assessment: Right knee pain, likely musculoskeletal strain vs early
osteoarthritis.

Plan: Ibuprofen for pain. Physical therapy referral. Imaging if no
improvement in 4 weeks.

Verdict: NO
Reasoning: Paraphrasing into clinical language ("erythema" for
"redness", "ROM" for "range of motion", "medially" for "the inside of
the knee", "no effusion" for "no swelling"). Assessment matches the
doctor's verbal differential. No substantive issue.

--- Example 3 (verdict: YES) ---

Transcript:
[doctor] Mrs. Allen, how have you been since starting the new
medication?
[patient] Pretty good, the headaches are mostly gone.
[doctor] Any side effects?
[patient] None that I've noticed.
[doctor] Let's continue the same plan. See you in three months.

Note:
Mrs. Allen reports significant improvement in headaches since starting
amitriptyline 50mg nightly. She denies side effects.

Plan: Continue amitriptyline 50mg HS. Return in three months.

Verdict: YES
Reasoning: Fabricated content. The transcript refers only to "the new
medication" -- no medication name, no dose, no schedule is mentioned.
"Amitriptyline 50mg nightly" has no transcript anchor. A clinician
reading this note would believe the patient is on a specific drug
that is in fact unspecified by the transcript.

--- Example 4 (verdict: YES) ---

Transcript:
[doctor] Any nausea, vomiting, or weight changes?
[patient] No nausea or vomiting. My weight has actually been stable.
[doctor] Let me check your abdomen. ... Soft, non-tender, no masses.

Note:
ROS: Patient reports recent 5-pound weight loss. Denies nausea or
vomiting.

Abdomen: Soft, non-tender, no masses appreciated.

Verdict: YES
```

```
Reasoning: Material misrepresentation. The patient said weight is
stable; the note asserts a 5-pound weight loss with a specific
number. Direction flipped and quantity fabricated. A clinician
reading this note would believe the patient is losing weight.

--- Example 5 (verdict: YES) ---

Transcript:
[doctor] How long have you had the chest pain?
[patient] A while now -- hard to say. Maybe a year?
[doctor] Any cardiac workup before?
[patient] No, never.
[doctor] Given the history, I want an EKG today and probably a stress
test.

Note:
Patient reports a 12-month history of stable angina. Cardiac workup
is being initiated; EKG and stress test ordered.

Assessment: Stable angina, awaiting workup.

Verdict: YES
Reasoning: Substantive over-reach. The doctor is initiating workup to
EVALUATE the chest pain; the diagnosis "stable angina" was not given
and is exactly what the workup is meant to establish. A clinician
reading this note would believe the patient has a confirmed diagnosis
of angina -- fundamentally different clinical understanding.

### Transcript
{transcript}

### Clinical note
{summary}

### Instructions
1. Read the note. For each item that's worded differently from the
   transcript, ask the fundamental question: would a clinician form a
   meaningfully different understanding from this note than the
   transcript supports?
2. If you find at least one item where the answer is YES, flag the
   doc YES.
3. If every item is paraphrase / abbreviation / standard write-up /
   reasonable contextual inference / light synthesis / omission, flag
   the doc NO.
4. End your response with exactly one line:
   `Verdict: YES` or `Verdict: NO`
```

I THE COMPARATOR LITERATURE IN FULL

Section 2 names four literatures and states what each contributes; this appendix carries the passages that section no longer prints at full resolution, and collects the published human-detection figures that Appendix L's human-baseline limitation summarises in a sentence. Nothing here is a new claim: every figure was verified against its primary source between 17 and 23 August 2026 and is quoted at the resolution the primary supports.

I.1 THE CONCURRENT AND ADJACENT STUDIES, IN FULL

The reviewing role, on non-clinical answer keys. Concurrent work locates the same asymmetry in the reviewing role itself: models asked to critique a key with planted defects catch inserted items six to seven times more often than omitted ones in that study's single-reviewer condition, and the authors argue the reason is structural - a reviewer can interrogate only candidates it can name, so

certifying that nothing is missing is itself an enumeration task (Chen et al., 2026). Their acceptable sets are finite and constructed and their omissions binary. Ours is the setting they leave untested: free clinical text with no enumerable answer set, omissions graded by severity and by surviving trace, and a source document on hand to supply the candidates their reviewer lacks.

Three further angles on judge reliability, including the one that points the other way. A perturbation study (Dahlberg et al., 2026) shows a clinical judge responding to deletions at roughly half its response to modifications, across three languages, without remarking on it. One mental-health study (Hussain et al., 2026) reports the opposite direction on prevalence-sensitive F1, a task and metric difference rather than a contradiction. Two concurrent vendor-authored studies pose the reliability question from adjacent angles: ten external clinicians agreed with each other only fairly on whether an automated judge's flag was genuine (AC1 0.24), with the judge inside that envelope (Bergman et al., 2026a), and re-running one hallucination judge under two criteria on the same 100 SOAP notes moved the measured rate from 35.2% to 9.1% (Vachhani et al., 2026) - a criterion effect on the commission side of the kind Section 6 measures on the omission side.

The instrument's own share of a published rate. A concurrent five-country paired study makes the instrument's share measurable from inside one vendor: on the same AI-generated notes, clinician adjudication found errors in 6.2% where a calibrated automated reviewer found 24.4%, and the authors publish both figures rather than choosing one (Bergman et al., 2026b).

I.2 Published human detection of omissions

We did not run a human baseline. The published record supplies the comparators: in a planted-error study, reviewing physicians caught omissions at roughly the same rate as objective errors, about a quarter to a third of each (Biro et al., 2025b) - no human asymmetry, at levels in the same range as our best configurations' 24.6 to 36.9%, though on a different task and corpus; MEDEC's two physician baselines reach 0.81 and 0.69 error-flag accuracy on commissions on its MS-test subset (document-level flags, not a per-error catch rate); clinicians reading with the source document alongside recall 50.4% of omissions in one review study (Bedi et al., 2026), an optimistic upper bound; and radiologists catch word-level, internally detectable omissions at rates from 42% to 92% (mean 61.8% across their twelve readers) (Shen et al., 2026), an easier task than semantic absence. No published figure covers human detection of omissions whose restatement survives elsewhere; if humans also miss those, that case is a document-design problem as much as a judge problem.

J The three judges on real vendor notes

Everything in Sections 5 to 7 is measured on constructed absence: pairs in which exactly one fact was removed by design. The companion census (Fox et al., 2026) supplies the other kind of evidence - real notes written by three deployed scribes, with omissions found and verified by an independent panel - and this appendix reports the one run in the study that points the paper's judges at them. Nothing was tuned on this corpus: each judge runs once, at the rule the benchmark chapters publish for it. Section 8's injected-errors limitation summarises the result; this is the full record.

J.1 Design

The set is **261 of the census's 565 notes**: all **87** that carry a panel-verified omission, and **174** of the 388 on which twelve discovery passes and a two-family verification panel found nothing at all, drawn stratified by product and corpus stratum (proportional, largest remainder, seed 20260825, worst marginal gap against the 388-note pool 0.32 percentage points). The 90 notes whose only verified findings are non-omission classes are excluded - they are neither clean nor omission-positive - and three of the 90, whose findings match no published class at all, fall under the same rule rather than being moved quietly into the clean class. A note's class is recomputed from its findings through the study's taxonomy code rather than read off a stored field: 155 of the census's 618 verified findings carry no stored top-tier class, and the stored field alone would miss 14 of the 87 omission notes.

Three judges run at their own published rules: the best monolithic design (the winsorised eight-sample mean, flag below 8.0), the evolved prompt (route two in the released data) at its deployed flag-below-10 rule, and the enumerate-then-check pipeline's three-stage tier (route one there) under

the per-fact critical rule. The pipeline must be the three-stage tier here: the two-stage tier judges the unaudited fact list, so its verdicts carry no severity grades and the critical rule cannot be stated on its output at all (the only rule it can express, any fact absent, fires on 72.3% of the benchmark's own clean twins). The census's 142 consultations are a strict superset of the evaluation set's 112, so extraction and audit artefacts are reused wherever they exist: of the 131 consultations these notes span, 103 had both cached and 28 were extracted and audited fresh. The run produced **783 judgements with zero parse failures**. Its records judge the withheld vendor notes and are handled under the same terms as the notes themselves (the ethics statement and Appendix F).

J.2 Every operating point moves, and none dominates

route	rule as published	detection, 87 verified-omission notes	false alarms, 174 no-finding notes	the same rule on the benchmark
monolithic (FC / score / k=8)	winsorised mean below 8.0	29.9% (26/87) [21.3, 40.2]	9.8% (17/174) [6.2, 15.1]	8.3% at 6.5%, swept (Section 5)
the pipeline under the per-fact rule (route one)	any critical fact absent	51.7% (45/87) [41.4, 61.9]	22.4% (39/174) [16.9, 29.2]	24.6% at 2.7% (Section 7)
the evolved prompt (route two)	score below 10	85.1% (74/87) [76.1, 91.1]	44.2% (77/174) [37.1, 51.7]	36.9% at 6.2% (Section 7)

Detection roughly doubles for both methods and more than triples for the monolithic judge, and false-alarm rates rise 8.3-fold on the pipeline, 7.1-fold on the evolved prompt and by half on the monolithic judge, so the three sit in a strict ordering in which whichever detects more also flags more clean notes. A threshold established on constructed clean twins does not carry to a production corpus - a caution about deployment rather than about the judges, and the reason the benchmark's operating points are quoted nowhere in this appendix as expectations for real traffic.

The mechanism is measurable from the stored per-fact verdicts. The pipeline checks the same number of audited facts per note on both corpora - 37.6 on the benchmark's held-out notes against 37.3 here - but it judges **2.0 facts absent per real note against 1.1 on the benchmark**, and its second look is called over 795 flagged facts against 299. A constructed pair has exactly one fact removed; a real note is simply less complete, which is also why the pipeline's check-and-second-look stage costs $0.0667 a real note against the benchmark's $0.0440 (the per-consultation extraction and audit stages are unchanged; the benchmark's full-pipeline price is $0.45). The "false alarm" column above should be read with that in mind: the 174 notes are clean only as far as the census's discovery reached, and the same verdicts say they are not fact-complete.

J.3 The evolved prompt's calibration, measured failing and re-established

Section 7 states that the evolved prompt's threshold is a calibration, not a design guarantee: it works because clean notes score exactly 10, which 96 to 98% of held-out twins and 90 to 94% of evaluation-set twins do. On real no-finding notes **97 of 174 score exactly 10 - 55.8%** - with 54 more at 9, so flag-below-10 fires on 44.2% of them. That is the predicted failure, measured. The signal survives re-calibration: swept on the same records, the evolved prompt reads **25.3% detection at 5.2% false alarms at flag-below-7** (32.2% at 5.2% at flag-below-8), while the monolithic judge has no comparable threshold - 9.2% at 1.1% at its own below-7, and 49.4% false alarms by below-9. So the ordering between those two survives the move to real errors even though the thresholds do not. The two are not read at the same false-alarm rate here, since the monolithic judge's threshold is the quieter of the pair, and no monolithic threshold reaches the evolved prompt's detection inside the 5% false-alarm band. (The per-fact rule carries no swept threshold and was not re-established.) Calibration also varies by product more than anything else measured here - the evolved prompt's false-alarm rate spans 30 to 71% across the three scribes - so a per-deployment clean set is not optional for this method, and per-product rates measured this way are properties of the instrument as much as of any product.

J.4 THE PIPELINE'S FLAGS NAME THE RIGHT FACT

A per-fact flag can be checked against the census's independently verified finding: two verbalisations of the same absence, one by the judge from the transcript, one by the panel. On the 87 verified-omission notes the rule fired on 45, naming 60 facts, and **34 of the 45 flagged notes (75.6% [61.3, 85.8]) name a fact that matches the panel's verified omission** under a lexical-containment screen at a 0.3 threshold (68.9% at 0.4, 57.8% at 0.5). The pairs are legible rather than marginal - "ulnar styloid fracture was stated to be present" against "documented ulnar styloid fracture finding omitted (and negated) in the note". This is an indicative screen, not an adjudication: read it as about three flags in four naming the fact the panel independently verified, not as a precision estimate with two significant figures. No injected benchmark can produce this number, because agreement there is true by construction.

J.5 WHAT THESE RATES ARE, AND ARE NOT

- **The false-alarm rates are upper bounds.** Discovery recall on the census is unmeasured, panel-cleared notes are not fact-complete (J.2), and the census's own standards experiment moves the share of notes carrying something reportable from 27.8% to 96.5% depending on the review instruction. A flag on a panel-cleared note is therefore not automatically wrong, and the three rates above are ceilings on the judges' error, not measurements of it.
- **Detection is conservative.** Ground truth is the panel's strict standard, so the 87-note positive class is a floor on the notes that carry an omission.
- **Two of the three flags are note-level.** The 87 notes carry 284 verified findings across all classes, so a monolithic or evolved-prompt flag cannot be attributed to the omission rather than a co-occurring failure; only the pipeline's flag names a fact, which is why J.4 exists for it alone.
- **The evolved prompt runs once here, against a deployed rule that is a majority of three.** A single draw detects and false-alarms more than the majority does (Section 7.4's estimator-matched readings), so its figures here sit at a noisier point than its deployed rule would.
- **The two "critical" gradings are different instruments** - the pipeline's audit grades on the benchmark's three-value rubric, the census on a two-value scale - so no severity-conditioned rate here is comparable with the benchmark's without naming the grader.
- **One run, one judge family, one corpus**, with no replicates on any arm; and these rates share no denominator with the benchmark's, so they are reported beside its tables and never inside them.

K THE EXPERIMENTAL PROGRAMME, THE JUDGE ROSTER, AND THE FULL ABLATION GRID

This appendix carries the maps, tables and figures the main text refers to but does not print: the programme end to end, which numbers live on which set of notes, the full roster of judges under test, every cell of the eight-design ablation, the per-lever attribution behind Section 6, the survey of every design on the detection-against-false-alarm plane, the per-fact verdicts read two ways, the practitioner tables, and the case where both methods fail.

K.1 THE PROGRAMME IN ONE LINE

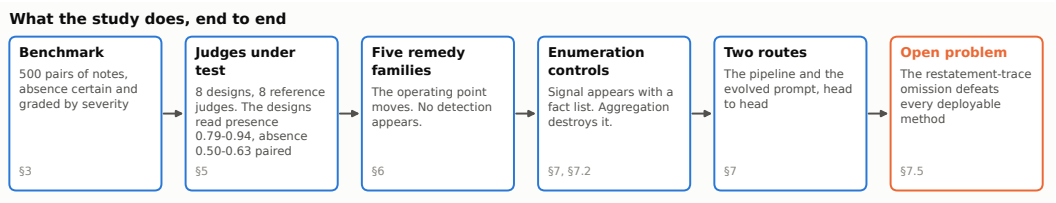


Figure K.1: The experimental programme in one line. The benchmark of Section 3 feeds an eight-design ablation and eight reference judges (Section 5, the asymmetry); five remedy families fail to restore single-note detection (Section 6); two enumeration controls locate the mechanism (Section 7.2); two methods - the enumerate-then-check pipeline and the evolved prompt - implement the conversion of absence into presence checks and are compared head to head (Section 7.4); and one case defeats every deployable method (Section 7.5). Counts, where a box carries them, are the pairs or designs involved. The box numbering is this paper's section numbering.

K.2 WHICH NUMBERS LIVE ON WHICH SET

Four sets of notes carry every judge number measured on this paper's own benchmark, and no rate is compared across two of them except where the sentence says so, at Appendix J's transfer measurement. Table K.1 is the map.

Table K.1: Which numbers live on which set. Every judge number measured on this paper's own benchmark comes from exactly one of these four sets, and rates measured on different sets are never compared or ranked against each other, with one deliberate exception, flagged in the sentence that makes it: Appendix J's transfer measurement, where the movement of an operating point between two sets is itself the result. The first three are built from the benchmark of Section 3; the fourth is real scribe output with the companion census's verified findings as ground truth.

Set	What it holds	Its role	The claims measured on it
Exploratory pass	38 pairs (32 of them omissions) against 32 clean notes, judged while the pipeline was still being designed	development readings only; ten of its consultations recur in the confirmation subset, with extraction caches reused	no headline claim: the pipeline's first paired read (0.828) and the per-fact rule's back-check (25.0% at 6.2%), quoted only as replication checks
Held-out confirmation subset	151 pairs over 47 consultations: 131 omissions, 20 commissions, 47 clean twins	committed before any optimisation began; never seen by any optimiser	the pipeline tiers' paired scores (three replicates), the per-fact rule's 20.6% at 2.1%, the reasoning-budget ladder, the optimiser's held-out tests, the second-family transfer. Marked † in Table K.5 and named per line in the figures
Evaluation set	495 pairs over 112 consultations: 293 omissions, 202 commissions, 112 clean twins (five optimiser-seen pairs excluded)	the full released benchmark; the eight designs' and the reference judges' flag rules fixed before the run, the two route rules established post hoc	the asymmetry across the eight designs (0.79–0.94 against 0.50–0.63), the reference judges, the remedy families, both routes' deployment rules (24.6% at 2.7%; 36.9% at 6.2%) and their head-to-head
Real vendor notes	261 of the companion census's 565 notes: 87 carrying a panel-verified omission, 174 with no verified finding	nothing tuned on it; each route runs once at its benchmark rule	whether the benchmark's operating points transfer to production notes (Section 7.4, Section 8, Appendix J)

K.3 THE JUDGES UNDER TEST

Table K.2: The eighteen judge configurations. All single-call judges run with no reasoning effort inside a 1,024-token answer cap; the multi-call recipes carry per-call budgets (Appendix D). "Evaluation set" is the 495 pairs and 112 clean twins; "held-out subset" the 151 pairs and 47 twins over 47 consultations that no optimiser ever saw. Six configurations are grouped as reference judges here; the figures plot those six together with the RAGAS-style recipe and the optimiser's winner, and call all eight the reference judges. Section 7 additionally re-runs a subset of these prompts with a reasoning budget; those re-runs are the same prompts, not new designs.

Group	Design	n	Answer returned	Scored on
The ablation	every combination of three choices: check faithfulness alone or faithfulness and completeness × answer yes/no or score 0–10 × ask once or eight times	8	a verdict or a score	evaluation set, 3 replicates
Reference judges	the deployed faithfulness judge, in three wordings differing only in how they treat omissions	3	a verdict	evaluation set
	G-Eval: scores several named dimensions, each as one open question	1	scores	evaluation set
	checklist judge: generates a per-consultation checklist (∼25 items) and answers each	1	item answers, read as a score	evaluation set
	the engineered omission judge: the strongest single-prompt judge we could write by hand, with eight few-shots	1	a score + free text	evaluation set
RAGAS-style recipe	per-fact coverage in the RAGAS lineage, our own deliberately naive implementation: extract the transcript's facts once (∼62), check each against the note	1	per-fact verdicts, read as a score	evaluation set
Enumerate-then-check pipeline	two-stage: extract facts from the transcript alone, then one closed present-or-absent check per fact	1	per-fact verdicts	held-out subset
	three-stage: adds a critic audit, severity grades, a full/partial/absent verdict and the decision rule	1	graded per-fact verdicts	held-out subset
Optimiser's winner	the exploratory GEPA winner at the standard settings (its reasoning-budget re-run is Section 7's evolved prompt)	1	a score	evaluation set
		18		

K.4 THE EIGHT DESIGNS, EVERY CELL

Table K.3 gives every cell of the ablation, including the two designs Section 5 does not name and the detection-versus-false-alarm z behind its "largest gap is 1.02 standard errors".

Table K.3: The eight judge designs, on the 495-pair evaluation set. Each design crosses what the judge is told to check (faithfulness only, F; faithfulness and completeness, FC), how it answers (a yes/no verdict, bin; a 0-to-10 score, score) and how many times it is asked (once, k1; eight times with the answers combined by a winsorised mean, k8). Paired discrimination is the tie-adjusted probability that an errored note scores below its own verified-clean twin, where 0.500 is a coin flip; it needs the twin, so it is a ceiling rather than a deployment measure. Detection and false alarms are single-note flag rates at each design's own rule, fixed before the run (a FAIL verdict for the single-sample yes/no designs, a winsorised mean of the eight encoded verdicts at or below 5.0 for the eight-sample yes/no designs, a score of 7 or below for the single-sample scored designs, and a winsorised mean below 8.0 for the eight-sample scored ones), over 879 omission records and 336 clean records: three replicates of 293 omission pairs and 112 clean twins, with 202 commission pairs carrying the commission column. z is the two-proportion statistic for a design's detection against its own false alarms, pooled over the three replicates; pooling counts three passes over the same notes as independent observations and inflates z, so the per-note-majority reading in Section 5 (1.2 to 1.7 standard errors for the two yes/no completeness designs) is the honest strength of that gap. Where the flag line was drawn is not what limits these designs. The best-separating cut the data itself suggests - the threshold that maximises Youden's J (detection minus false alarms) on the mean-over-runs aggregate - lands at 8.17 (F-score-k1), 8.40 (F-score-k8), 8.50 (FC-score-k1) and 8.19 (FC-score-k8), every one of them above the 7 and 8.0 fixed before the run rather than below; and threshold-free, the area under the curve stays between 0.503 and 0.575 across all eight. A dash means the quantity is not swept: a single yes/no verdict leaves no score to sweep, and the eight-verdict aggregate is a vote fraction rather than a score. The AUC printed for a single-sample yes/no design is the area its single operating point defines; for the eight-sample yes/no designs it is computed over the eight-verdict vote fraction.

	paired discrimination		at its own flag rule				
Design	commissions	omissions	detection	false alarms	z	AUC, omissions	swept det. at FA $\leq$ 10%
Faithfulness only, yes/no, one sample (F-bin-k1)	0.875	0.500	9.6%	8.9%	0.34	0.503	–
Faithfulness only, yes/no, eight samples (F-bin-k8)	0.904	0.509	9.6%	7.4%	1.15	0.512	–
Faithfulness and completeness, yes/no, one sample (FC-bin-k1)	0.792	0.539	40.3%	31.8%	2.71	0.542	–
Faithfulness and completeness, yes/no, eight samples (FC-bin-k8)	0.869	0.591	42.3%	32.1%	3.25	0.562	–
Faithfulness only, score, one sample (F-score-k1)	0.911	0.549	1.0%	1.5%	-0.68	0.549	1.0% @ 1.5%
Faithfulness only, score, eight samples (F-score-k8)	0.944	0.582	2.4%	2.1%	0.32	0.549	2.4% @ 2.1%
Faithfulness and completeness, score, one sample (FC-score-k1)	0.895	0.585	4.7%	5.7%	-0.71	0.573	4.7% @ 5.7%
Faithfulness and completeness, score, eight samples (FC-score-k8)	0.939	0.634	8.3%	6.5%	1.02	0.575	8.3% @ 6.5%

K.5 WHERE EACH REMEDY MOVES THE OPERATING POINT

Table K.4: What each lever bought, and what it did to the judge's operating point. Paired discrimination is the tie-adjusted probability that an errored note scores below its own verified-clean twin, so 0.500 is a coin flip and a lever worth $+0.04$ buys four correctly ordered pairs in every hundred, only where a clean twin exists to compare against. The first three levers are quoted as the span of four per-design deltas, each holding the other two choices fixed, because a single headline delta would hide how far the spans overlap: criterion and format are indistinguishable on the mean ($+0.052$ against $+0.053$) and ensembling is smallest ($+0.036$). The reading rule this table carries: strict win rates must never be quoted across designs that differ in sample count, because a finer aggregate splits ties and inflates the count. The completeness-scored pair is the case in point. Asking eight times rather than once takes the strict win rate on omissions from 27% to 43%, but over the same 293 pairs the tie rate falls from 63% to 41%, and once half the tie mass is counted the paired figure moves 0.585 to 0.634. That is a lever worth $+0.049$, not sixteen points. Every paired figure in this paper is tie-adjusted for that reason.

Lever	Paired delta on omissions	What it did to the operating point	Measured on
What the judge is told to check: faithfulness only, to faithfulness and completeness	+0.036 to +0.082 (mean +0.052) *four per-design deltas, holding format and sample count fixed*	Detection rises in all four comparisons, from 1.0-9.6% before to 4.7-42.3% after, and false alarms rise with it, 1.5-8.9% to 5.7-32.1%; of the eight designs only the two yes/no completeness ones have detection that clears their own false alarms on pooled judgements (z 2.71 and 3.25; 1.24 and 1.65 once repeat runs collapse to a per-note majority, the paper's unit of inference), and they do it at 31.8% and 32.1% false alarms.	the eight designs, 495-pair evaluation set, 3 replicates (879 omission records, 336 clean records)
How it answers: a yes/no verdict, to a 0-to-10 score	+0.043 to +0.073 (mean +0.053) *four per-design deltas, holding criterion and sample count fixed*	Both rates fall together, detection from 9.6-42.3% to 1.0-8.3% and false alarms from 7.4-32.1% to 1.5-6.5%, so the scored designs are quieter rather than sharper: their detection never clears their own false alarms by more than 1.02 standard errors.	the eight designs, 495-pair evaluation set, 3 replicates (879 omission records, 336 clean records)
How many times it is asked: once, to eight times combined by a winsorised mean	+0.009 to +0.052 (mean +0.036) *four per-design deltas, holding criterion and format fixed*	The operating point barely moves: the largest change is the completeness-scored pair, detection 4.7% to 8.3% at false alarms 5.7% to 6.5%, and no scored design's detection separates from its own false alarms at either sample count.	the eight designs, 495-pair evaluation set, 3 replicates (879 omission records, 336 clean records)
How the prompt words the omission instruction: the deployed faithfulness judge in three wordings, as shipped, with its omissions-are-not-errors exclusion deleted, and with an affirmative instruction to report omissions added	+0.051 (one comparison, not a span) *the affirmative instruction against the same prompt with the exclusion merely deleted; paired 0.518 as shipped, 0.505 deleted, 0.556 affirmative*	The affirmative instruction takes detection from 13.1% to 25.6% (77 and 150 of 586 omission judgements, against 16.4% as shipped) while false alarms hold at 27 of 224 clean judgements (12.1%) for all three wordings, the same rate though not the same notes; deleting the exclusion on its own takes paired discrimination from 0.518 to 0.505 and detection from 16.4% to 13.1%.	three wordings of the deployed judge, 495-pair evaluation set, 2 replicates (586 omission records, 224 clean records)
Letting an optimiser rewrite the prompt: reflective prompt evolution, aimed at the per-fact checker and at the best monolithic judge	+0.052 and +0.073 on the acceptance sets, -0.023 and +0.004 held out *two search legs, each measured on the 20-consultation acceptance set that chose the winner and then once on the 47-consultation held-out set*	Held out, the per-fact checker's winner gains 0.8 points of detection (22 against 21 of 131 omissions) at the same 8.5% false alarms, and the monolithic winner loses 4.6 points (11 against 17 of 131) at the same 6.4%. The earlier campaign's winner carried a false-alarm rate of 9.1% (2 of 22 clean development notes) to 21.9% (49 of 224) on data it had never seen, at 23.0% detection (135 of 586).	acceptance set 48 omission pairs and 20 clean twins; held-out set 131 omission pairs and 47 clean twins, one replicate; the earlier campaign's winner on the 495-pair evaluation set, 2 replicates

K.6 EVERY JUDGE DESIGN ON THE DETECTION-AGAINST-FALSE-ALARM PLANE

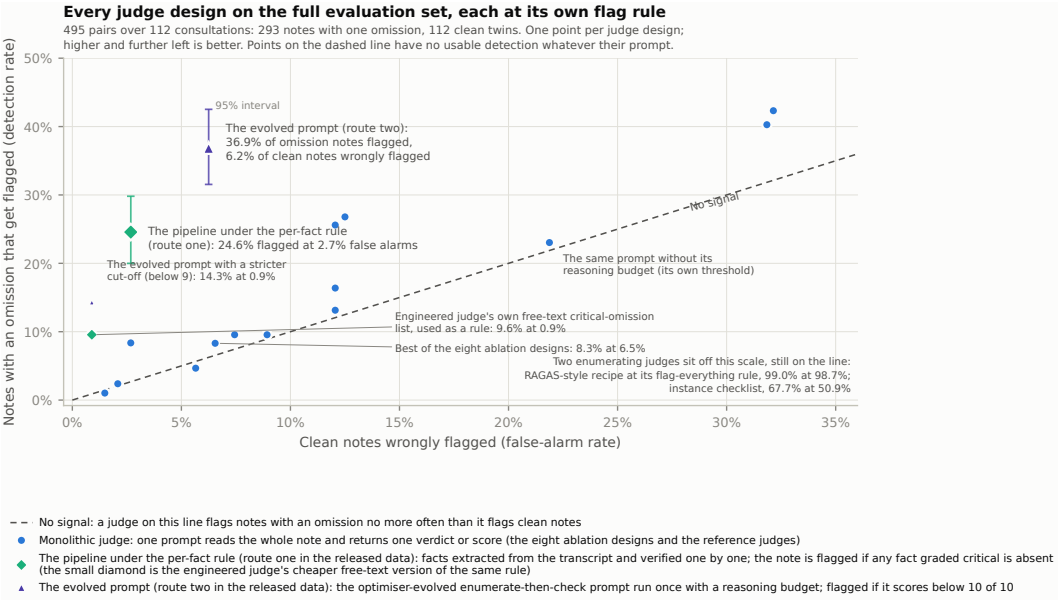


Figure K.2: Every judge design at its own flag rule: the levers slide a point along the no-signal line, and only the two methods leave it. Across, how often a judge wrongly flags a clean note; up, how often it flags a note that really carries an omission. A useful detector sits high and to the left; the dashed diagonal is no signal, a judge that flags both kinds of note equally often. The legend defines the kinds of point. Every design is on the 495-pair evaluation set (293 notes with one omission, 112 clean twins), each at its own flag rule - defined in Section 4, fixed before the run for the eight designs and the reference judges, established post hoc for the two methods' rules (Sections 7.3 and 7.4). The eight designs and the deployed-judge wordings sit on or beside the line, or leave it only at false-alarm rates well above the 10% a deployment could carry, and the levers of Section 6 only slide a point along it. Two points sit well clear of it, each with a Wilson 95% whisker on its detection rate: the pipeline under the per-fact rule at 24.6% detection on 2.7% false alarms, and the evolved prompt at 36.9% on 6.2%. Four further points are labelled for reference: the evolved prompt with a stricter cut-off, the same prompt run without its reasoning budget, the engineered judge's own free-text critical-omission list read as a rule, and the best of the eight ablation designs at 8.3% on 6.5%. Two enumerating judges sit off the scale and are stated in the panel - the RAGAS-style recipe at its flag-everything rule and the instance checklist. "Route one" and "route two" in the panel are the two methods' handles in the released data. Every unlabelled point's values are in Table K.3, Table K.4 or Table K.5; how the pipeline's flag is made from its verdicts is Figure K.3.

K.7 THE SAME PER-FACT VERDICTS, READ TWO WAYS

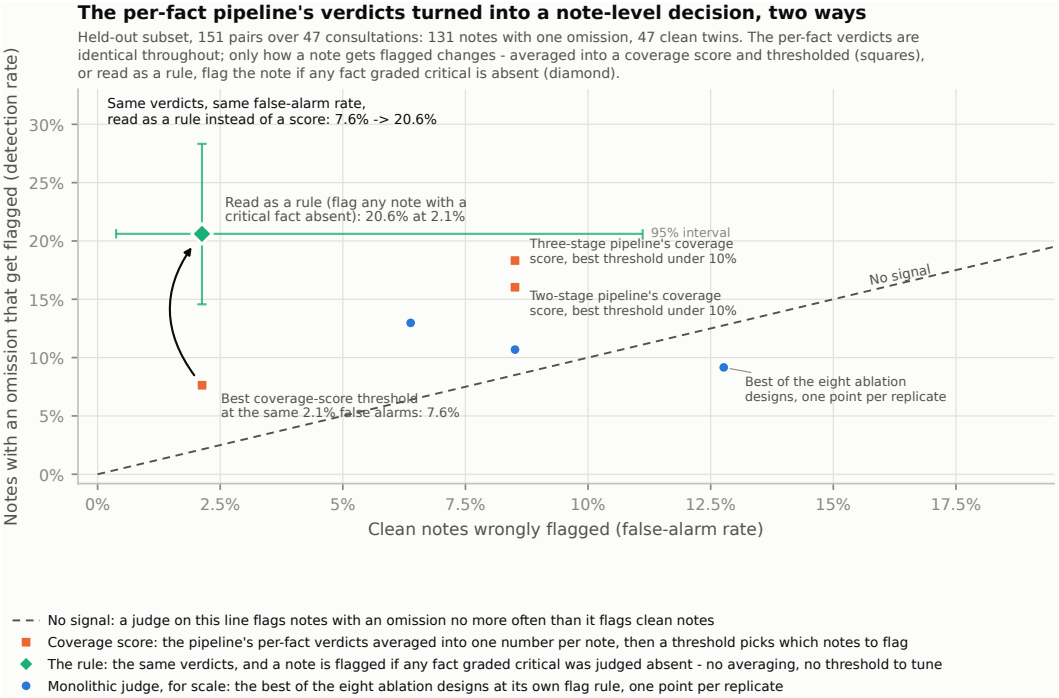


Figure K.3: The same per-fact verdicts, turned into a note-level decision two ways. A pipeline that returns one verdict per fact still has to decide, per note, whether to flag it. On the 151-pair held-out subset (131 notes with one omission, 47 clean twins) the three-stage pipeline's verdicts are read two ways with nothing else changed. Averaged into a coverage score and thresholded at the rule's own 2.1% false-alarm rate, they detect 7.6% of the notes with an omission; the best threshold inside a 10% false-alarm budget reaches 18.3% at 8.5% for the three-stage tier and 16.0% at 8.5% for the two-stage tier. Read as the rule - flag the note if any fact graded critical is judged absent - the identical verdicts detect 20.6% at 2.1% (the plotted rule point is replicate 1 of three; across the three the detection reads 20.6, 22.9 and 21.4%, with the false-alarm rate identical in every one). The best of the eight ablation designs is drawn for scale at its own rule, one point per replicate. Same verdicts, same noise, 2.7 times the detection: the decision is made on the verdicts, not on a score.

K.8 WHAT EACH RUNG DELIVERS

Table K.5: What each rung delivers. Reading the columns: paired discrimination as Section 4 defines it (0.500 is chance, and it is a ceiling - production has no clean twin); detection and false alarms on single notes, each design measured at its own flag rule (also defined there; fixed before the run except the two routes', Sections 7.3 and 7.4); and the best detection a design reaches with false alarms capped at 10%. Unmarked rows are the evaluation set - 293 omission pairs, 202 commission pairs and 112 clean twins - with the eight designs at three replicates, the reference judges at two, the per-fact rule at one and the evolved prompt's flag rates a majority over three runs, while † marks the 151-pair held-out subset (131 omission pairs, 20 commission pairs, 47 clean twins), whose rows are not comparable to the others. Four special cases. The two RAGAS-style rows' paired columns rest on 289 and 200 pairs, because parse failures cost that arm 14 of its 1,214 records; their absolute columns keep the full 586 omission and 224 clean records, an unparseable record counting as no flag. The coverage-score row repeats the recipe's three flag-rate cells, having no flag rule of its own. The two pipeline-tier rows print three-replicate paired means (Section 7.1 gives the two-stage tier's spread; Appendix L gives both), with their last column at replicate 1. The pipeline's deployable row prints its rule's flag rates, while its paired columns read the three-stage tier's underlying coverage score, one replicate - the rule itself is a yes/no flag with no score to rank pairs by. A dash means the quantity does not exist for that design - a yes/no verdict leaves no score to sweep, the free-text rival is a rule over a list, and the pipeline tiers write no flag of their own, so they are swept rather than read at a rule; the two route rules already sit inside the 10% cap, so their last column repeats their detection. Neither route dominates: the evolved prompt buys the higher-detection point at 2.3 times the rule's false-alarm rate and a tenth of its price, resting on a per-deployment calibration, while the rule buys the quiet point and each of its flags names the missing fact with an audit-stage severity grade - a grade the study's one clinician rater read about one grade above their own on routine content (Section 7.3 reads that lean as the rater's rather than the rubric's). What each row costs and what its flag contains are in Table K.6, over the same rows in the same order.

	Paired discrimination		At its own flag rule		
Approach	Omissions	Commissions	Detection	False alarms	Best detection at ≤10% false alarms
One call at standard settings, asked an open question					
Best of the eight grid designs (FC / score / k=8)	0.634	0.939	8.3%	6.5%	8.3%
Cheapest of the eight (F / score / k=1)	0.549	0.911	1.0%	1.5%	1.0%
Best yes/no design (FC / binary / k=8)	0.591	0.869	42.3%	32.1%	–
Engineered omission judge (hand-written, 8 few-shots)	0.639	0.754	8.4%	2.7%	8.4%
Deployed faithfulness judge, as shipped	0.518	0.886	16.4%	12.1%	–
Optimiser's exploratory winner, standard settings	0.549	0.861	23.0%	21.9%	7.7%
G-Eval (decomposes into dimensions, never enumerates)	0.568	0.890	26.8%	12.5%	–
Enumerate the transcript, read the result as a score					
Instance checklist (Zhou et al., 2025)	0.647	0.744	67.7%	50.9%	16.0%
RAGAS-style recipe (ours)	0.809	0.712	99.0%	98.7%	20.9%
Its coverage score alone	0.817	0.665	99.0%	98.7%	20.9%
Enumerate + check (pipeline B2)†	0.786	0.650	–	–	16.0%
Enumerate + audit + check (pipeline B3)†	0.762	0.617	–	–	18.3%
The deployable operating points, each at its own rule					
Enumerate-then-check pipeline (route one) - flag any critical fact judged absent	0.795	0.661	24.6%	2.7%	24.6%
Evolved prompt at its reasoning budget (route two, one call)	0.670	0.978	36.9%	6.2%	36.9%
Evolved prompt on the second judge family (gemini-3.1-pro, same prompt)†	0.659	0.975	34.4%	6.4%	34.4%
Free-text rival: the engineered judge's own critical-omission list	0.639	0.754	9.6%	0.9%	–

K.9 PRICES AND WHAT EACH FLAG CONTAINS

Table K.5 carries performance; Table K.6 carries the prices and the flag contents over the same rows in the same order, at the pinned credit rates whose measurement dates are in Appendix D.

Table K.6: What each rung costs, and what its flag contains. The rows and their order are those of Table K.5, which carries the same approaches' discrimination and operating points. Cost is measured from the run receipts in a production framing of one note per consultation, so per-consultation stages are not amortised across a benchmark's twins; amortised across this benchmark's twins the two pipeline tiers read $0.028 and $0.140 a note (the two ratios differ because the tiers spend different shares of their price on per-consultation stages). The difference between the two tiers' prices exceeds the audit stage's own 71% share of $0.45 because the three-stage tier also runs its closed check at dearer settings and a second look over flagged facts. What a flag carries is what the design hands the reader when it fires: a bare score or verdict, or the named missing fact, with or without a severity grade and a supporting quote. Hidden cost is what the row asks for beyond its ledgered price. ‡ marks designs that also generate one cached artefact per consultation whose cost is not separately ledgered, making those figures lower bounds; the two rows that read a threshold against a clean-note calibration need a labelled clean set per deployment to place that threshold, which no price here includes. A dash means the row asks for nothing beyond its price.

Approach	$/note	What a flag carries	Hidden cost
One call at standard settings, asked an open question			
Best of the eight grid designs (FC / score / k=8)	$0.036	A score	–
Cheapest of the eight (F / score / k=1)	$0.004	A score	–
Best yes/no design (FC / binary / k=8)	$0.035	A verdict	–
Engineered omission judge (hand-written, 8 few-shots)	$0.005	A score + free text	–
Deployed faithfulness judge, as shipped	$0.005	A verdict	–
Optimiser's exploratory winner, standard settings	$0.008	A score	–
G-Eval (decomposes into dimensions, never enumerates)	$0.008	A score	–
Enumerate the transcript, read the result as a score			
Instance checklist (Zhou et al., 2025)‡	$0.006	A score	One cached artefact per consultation, not ledgered
RAGAS-style recipe (ours)‡	$0.030	A score	One cached artefact per consultation, not ledgered
Its coverage score alone‡	$0.030	A score	One cached artefact per consultation, not ledgered
Enumerate + check (pipeline B2)	$0.094	A coverage score	–
Enumerate + audit + check (pipeline B3)	$0.45	A coverage score	–
The deployable operating points, each at its own rule			
Enumerate-then-check pipeline (route one) - flag any critical fact judged absent	$0.45	The named fact + severity	–
Evolved prompt at its reasoning budget (route two, one call)	$0.046	A score below 10	A clean-note calibration, re-established per deployment
Evolved prompt on the second judge family (gemini-3.1-pro, same prompt)	$0.028	A score below 10	A clean-note calibration, re-established per deployment
Free-text rival: the engineered judge's own critical-omission list	$0.005	The fact, free text	–

K.10 WHERE BOTH METHODS FAIL

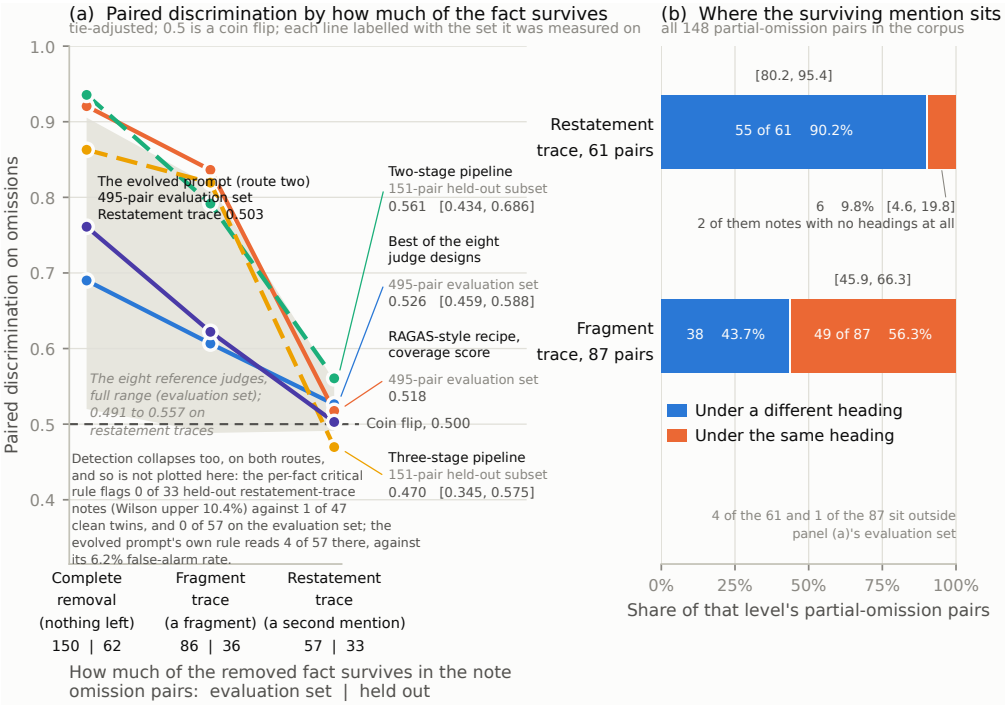


Figure K.4: Every method we measured in the primary judge family collapses on the omission whose restatement survives elsewhere in the note, and the supply of those cases says why. (a) Paired discrimination on omissions - how often a note carrying an omission scores below its own verified-clean twin, tie-adjusted, where 0.5 is a coin flip - split by how much of the removed fact survives: nothing (complete removal), a fragment of it (fragment trace), or an explicit or closely paraphrased second mention (restatement trace). Each line names the set it was measured on, and the two sets are read separately rather than ranked against each other: the best of the eight judge designs, the range across the eight reference judges, the RAGAS-style recipe (our implementation, read as its coverage score rather than at its own flag-everything rule), and the evolved prompt (route two in the released data, the optimiser's winner at reasoning effort high, three replicates averaged) are measured on the 495-pair evaluation set, whose 293 omission pairs split 150 / 86 / 57 across the three levels; the two-stage and three-stage pipelines are measured on the 151-pair held-out subset, whose 131 omission pairs split 62 / 36 / 33. The band spans all eight reference judges, the coverage recipe among them at the aggregate score it ships, so that recipe appears twice: once inside the band and once as its own line at the coverage component alone. The coverage recipe is also the one line with denominators of its own, 148 / 84 / 57, because parse failures cost it 14 of its 1,214 records. The interval printed beside each restatement-trace value is a consultation-clustered bootstrap 95% interval, computed over 35 consultations for the evaluation-set line that carries one and 20 for the held-out ones, and every one of them contains 0.5; the coverage recipe's and the evolved prompt's restatement-trace values have no such interval in the released artefacts and are quoted as point estimates. Detection collapses with discrimination, and is annotated rather than plotted because it is a rate at a rule rather than a discrimination score: the per-fact critical rule flags 0 of 33 held-out restatement-trace notes against 1 of 47 clean twins, 0 of 57 on the full evaluation set against 3 of 112, and the evolved prompt's own rule reads 4 of 57 there against its 6.2% false-alarm rate. (b) Where the surviving mention sits, over all 148 partial-omission pairs in the corpus (61 restatement-trace, of which 57 are in the evaluation set, and 87 fragment-trace, of which 86 are): in 55 of 61 restatement-trace pairs, 90.2% [80.2, 95.4], the survivor sits under a different heading from the instance that was removed, while 49 of 87 fragment-trace pairs, 56.3% [45.9, 66.3], keep it under the same heading; two of the six same-heading restatement traces are notes with no headings at all. Reading: once a restatement trace survives, every method plotted here sits at or below its own noise floor, and the second judge family's checker, which is not plotted, is the one exception the study holds: it discriminates the class at 0.636 paired while the rule still catches none (Section 7.5). The supply explains why - a restatement trace is almost always a different-heading survivor, recoverable only by reading the note end to end.

L LIMITATIONS IN FULL

Section 8 states the four that bound the results directly. This appendix is the full list.

No external anchor for the omission half. The principal unresolved validity question is whether the transcript-derived omission labels correspond to independent clinician judgements of documentation failure. The commission arm has an external anchor in MEDEC; the omission arm has none, because the only external physician-labelled benchmark whose task matches ours is drawn entirely from a corpus that is one of our own strata. Until such an anchor exists, the omission results are internal to the instrument this paper builds: constructed, verified and clinician-checked, but not independently labelled. Establishing that correspondence is the first thing future work has to do.

One judge model family for the headline results. Every headline judgement in Sections 5 to 7 comes from one judge model family. The circularity Section 7.1 answers also has a model-level form that the three controls there do not fully close: the pipeline's extraction stage runs on the model that built the fact sheets and its audit on the model that verified the pairs (Appendix D), so construction and judging share model families as well as a task shape. The RAGAS-style recipe is the control that does not, extracting its facts with the judge model rather than with the model that built the answer key. The second-family checks reproduce the asymmetry, the restatement-trace collapse (with Section 7.5's nuance: the transferred checker discriminates the class while the rule still catches none) and the pipeline's lead (Appendix G); the evolved prompt's calibration transfers too (Section 7.4). Single-note detection proves judge-dependent, so single-note claims are scoped to the family measured. One confound is disclosed: reasoning effort is not matched tier-for-tier across the two families in the Section 5 checks, because the second family's endpoint refuses a no-reasoning setting; the evolved-prompt transfer does not carry that confound, since both families run at high effort.

What the compute settings bound. Sections 5 and 6 are measured with the judge given no reasoning effort inside a 1,024-token cap, and all three prompt searches executed their candidates the same way, so the search explored a compute regime a deployment need not live in. Section 7 measures what a reasoning budget does to those prompts, but only to prompts already selected; a search re-run at a larger budget is a different experiment and we did not run it. One control is missing on the same subset: at matched compute the plain completeness-scored prompt matches the evolved prompt on paired discrimination (0.670 against 0.669), so the evolved prompt's measured advantage is the clean-note calibration and the operating point it admits, and whether a plain prompt given the same budget would score clean notes at 10 and so admit a flag rule of its own was not tested. Those re-runs add one limit of their own: the evolved prompt's clean-note calibration is a behaviour of this corpus and judge, not a property of the design.

What the head-to-head can and cannot say. The Section 7 comparison is at each method's own deployment threshold, not at matched false-alarm rates; its estimator asymmetry is measured in Section 7.4, and removing it strengthens the contrast rather than weakening it. The evaluation-set extension was run without an analysis plan fixed in advance: the tests reported are the ones established on the held-out sample, applied unchanged, with no new test chosen after seeing the data. The per-fact rule still has one replicate on the evaluation set - only the held-out subset carries three - so the equalisation reads the evolved prompt down rather than replicating the rule up.

One clinician on this paper's own items, and the difference between endorsement and blinding. A physician author validated 94 items across two sittings - 70 across six stages of the benchmark and pipeline, and 24 more on the audit stage's own severity grades. Stages where the item itself had to state the verdict being audited are **endorsements** - a clinician checked and did not object; only the panel-cut, severity and judge-disagreement stages were structurally blinded, and only those are measurements. They are an author who designed the instruments they graded, and there was one rater across both sittings with no repeated item, so no within-rater figure exists. Every stage holds 6 to 20 items, so four stages reading 100% with lower bounds of 61–72% rule out gross failure and nothing finer: the intervals are the result, the point estimates are not. Severity-conditioned results use the benchmark's grades; the per-fact rule fires on the audit stage's own. One thing has now been checked from outside: the severity rubric. An independent clinician graded against it, blind, in the companion census (Fox et al., 2026) (Appendix A). That is external evidence the rubric reproduces to within a grade, and it is also why the one-grade lean reported in Section 7.3 is now read as a property of the rater rather than of the rubric - the two clinicians lean opposite ways. It is not a second reading of this paper's own items: nobody outside the author group has adjudicated a judge output here, and that gap is the one this limitation names.

Injected, not naturally occurring, errors. The companion census (Fox et al., 2026) shows the classes are real, but injected pairs are cleaner than production failures, and we measured the gap on both arms. On commissions, the same judge scores paired 0.827 on MEDEC's physician-curated errors against 0.939 on our constructed ones. On omissions, Section 7.4's run against real vendor notes shows every operating point moving - detection and false alarms both rise steeply at the benchmark's own rules - so the benchmark's operating points do not transfer to production traffic, though once thresholds are re-established there the evolved prompt still detects more than the monolithic judge at half its false-alarm rate (Appendix J). The false-alarm reads there are upper bounds, not estimates: the census's cleared notes are clean only to the strictness of its discovery, and its own standards experiment moves the share of notes carrying something reportable from 27.8% to 96.5%. All commission pairs also come from the earlier build, so the construction-cohort robustness slice exists only for omissions, and a build effect acting on both classes at once cannot be excluded by slicing. Each pair also carries exactly one planted error where a real note carries several (the pipeline judges 2.0 facts absent per real note against 1.1 on the benchmark's pairs), so nothing here measures how a judge behaves when failures interact. One further bound runs through the constructed side: the clean twins are verified complete against the audited fact sheet rather than against the transcript, so every clean-note false-alarm rate this paper reports for an enumerating judge is an upper bound.

What the held-out subset is, and is not, held out from. Ten of the confirmation's 47 consultations also appear in the exploratory pass with extraction caches reused, so it is a fully held-out test of the judging and only a partial one of the extraction. The pipeline tiers' paired scores and the per-fact rule carry three runs each, with three-run spreads of 0.031 for the two-stage tier and 0.015 for the three-stage, against the monolithic baseline's 0.118, a quarter to an eighth; the swept single-note figures and the per-trace splits remain single runs. The subset is held out with respect to optimisation, not to the number of analyses run over it - the reasoning-budget re-runs, fact-list ablation and family transfer all read the same 151 pairs, with no multiplicity correction - and the per-fact rule itself was formulated post hoc, then checked on the earlier exploratory pass; the evolved prompt's flag-below-10 threshold was likewise read off the first replicate's scores and then held unchanged.

Human detection, from the literature rather than our own measurement. We did not run a human baseline. The published comparators run from about 25% to 92%, depending on the task and on whether the reader had the source document alongside. Our best configurations detect 24.6 to 36.9%, the same range as the one comparator in which physicians reviewed without the source document - they caught about a quarter to a third of planted omissions - while the higher human figures come from easier tasks. None of this settles the question. Appendix I collects every figure we could find, and none of them covers the restatement case.

Corpus limits. English-language UK and US primary and ambulatory care only, and no note carries an EHR structured field, so we cannot speak to structured-field omissions at all. Two cost levers for the pipeline, folding severity grading into the extraction call and a cheaper extraction model, are identified rather than tested.