MULTI-COUNTRY PAIRED SIMULATION · WITH A VALIDATED ERROR-DETECTION INSTRUMENT

Quality, consistency and clinical safety of AI-generated versus clinician-written clinical notes

Clinical AI Research · **study sponsor**

Henry Bergman, Ben Austin, Sarah Ali, Elliot Taylor, Merle Fiedler, Celene Sandiford, Rebecca Blanchard, Carlos Casanovas, Victor Najar, Giulia Pedrazzini, Theo Markopouliotis, Ferdinand Vermersch · validation arm additionally **Rohan Sanghera**. All: Clinical AI Research, Heidi Health. Presenting author affiliation: NHS England Clinical AI Fellowship.

Presenting body · NHS England Clinical AI Fellowship

+5.08
PDQI-9 **quality advantage**, AI over clinician (40.6 vs 35.6; d_z = 0.55)

~5×
fewer "poor" notes from the AI (5.7% vs 27.8% scored < 32)

68.1%
of consultations where the **AI note was clinically dominant**

7.3×
more **severe omissions** in clinician notes

2.4×
more **severe hallucinations** in clinician notes

Background

Ambient AI scribes are deploying at scale, yet the comparative evidence is dominated by single-site, single-language studies that rely on **insensitive human review** to find errors. The decisive question is not whether AI notes are perfect, but whether they are **safer and better than the clinician notes they replace** — judged on the same consultation, against the junior-to-middle-grade clinicians who write most routine documentation.

Methods

Paired simulation · 5 countries

English/UK, Spanish, Italian, French, German. **385 paired consultations, 770 notes.** Each consultation documented independently by the AI scribe (Heidi — unedited first draft, a pre-specified conservative choice) and a junior-to-middle-grade clinician (~40 writers, the primary documentation workforce).

Known ground truth

Actor-performed scripted consultations seeded with naturalistic dysfluency. **4 specialties** (ED, GP, Medical, Surgical) × **4 acoustic conditions** (clean, 40/50/60 dB noise).

Quality & safety

Quality on PDQI-9 (37 blinded evaluators). Errors — unsupported statements / "hallucinations" and omissions — identified under **two detection arms**: clinician adjudication (false-positive-stripped) and a calibrated automated reviewer. Every error risk-graded (severity × likelihood, MHRA-aligned) by an independent three-model panel (GPT-5.5, Claude Opus 4.8, Gemini 2.5 Pro).

Analysis

Pre-specified before pooling (POOLED_SAP.md). Paired Wilcoxon primary; Bayesian crossed-random-effects model. Co-primaries = PDQI-9 quality + Critical+High error burden under both arms.

Head-to-head, per consultation

Which note was better of the pair?

Combining quality & severe-error burden, the AI note was **clinically dominant in 68.1%** of consultations (sign test $p=2.6\times10^{-17}$).

AI scribe **Clinician** — Consistent across all panels

STUDY AT A GLANCE

5 countries & languages	385 paired consultations
770 clinical notes adjudicated	37 blinded evaluators

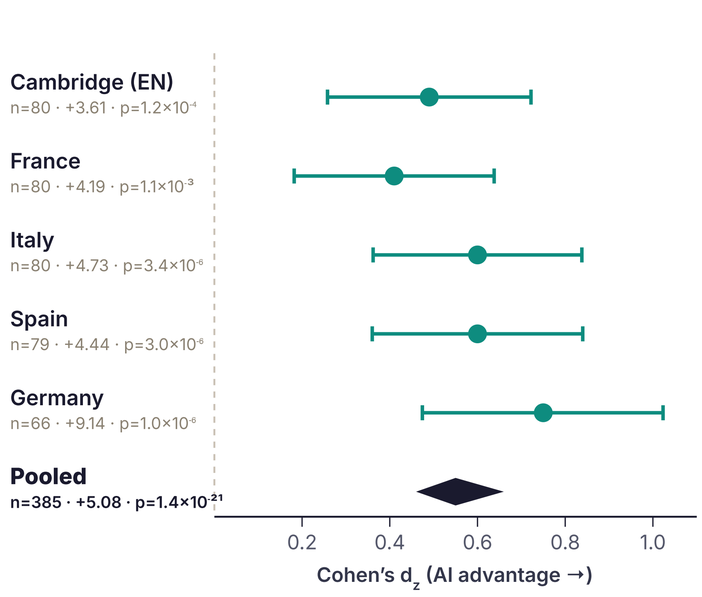
Results

The AI note is higher-quality, more consistent and lower-risk — in five languages, under two error detectors

FIG A · FOREST PLOT

The advantage replicates in all five languages

Per-site Cohen's d_z (AI − clinician) with 95% CI; pooled diamond d_z 0.55 (0.46–0.66), Wilcoxon $p=1.4\times10^{-21}$.

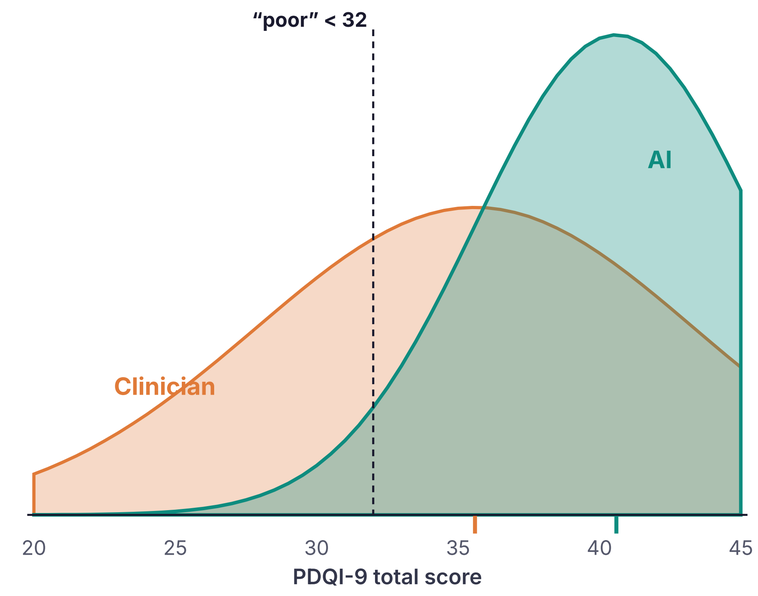


Per-site 95% CI from the paired-d_z SE approximation (reproduces the reported pooled CI). I^2 = 62.6%.

FIG B · CONSISTENCY

The AI rarely writes a poor note

Paired PDQI-9 distributions. SD 4.97 vs 7.76 (Levene $p=4.2\times10^{-13}$); 5.7% vs 27.8% below 32.

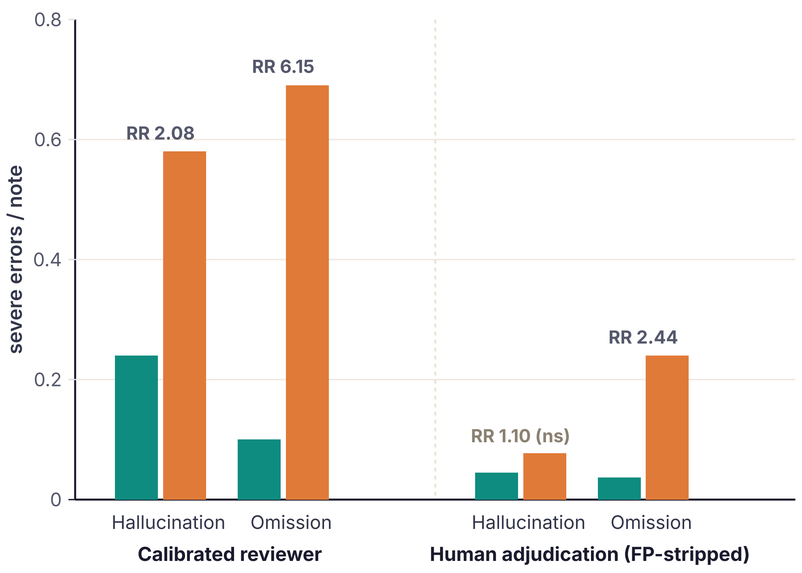


66% vs 37% score ≥40. Bayesian clinician deficit −5.11 (94% HDI −5.88 to −4.37; posterior P(clinician < AI) > 0.999). Evaluator variance (SD 3.65) dwarfs writer (0.81) and case (0.39) — the rationale for pairing.

FIG C · SAFETY

Fewer severe errors — direction is detector-independent

Severe (Critical+High) errors per note, by type and detection arm. The magnitude, not the direction, depends on the detector.

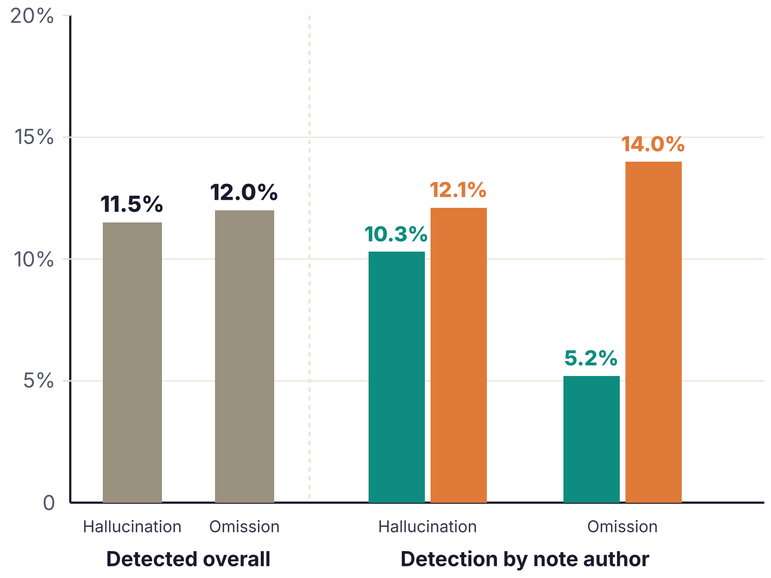


Calibrated: hallucination $p=1.6\times10^{-9}$, omission $p=8.8\times10^{-26}$. Failure modes differ — the AI occasionally fabricates; the clinician more often **omits** (medications, safety-netting, counselling). **Critical-tier: 19.5% of clinician vs 8.2% of AI notes.**

FIG D · DETECTION SENSITIVITY

Human review misses most errors — most of all on AI notes

% of automated-reviewer-confirmed errors that clinician reviewers detected.



Fluent AI prose hides its residual errors from human oversight — reviewers were least sensitive to omissions on AI notes (5.2% vs 14.0%; hallucinations 10.3% vs 12.1%). An argument for automated, in-workflow verification over unaided human review.

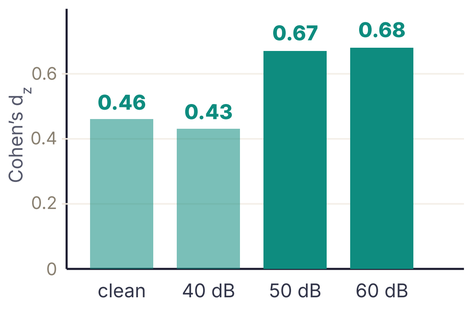
PDQI-9 ITEM PROFILE

Gains largest on completeness & accuracy

All 9 items favour AI, significant after FDR (4 shown).

NOISE DOSE-RESPONSE

Noisier audio, larger AI edge



Bars = Cohen's d_z, the standardised paired PDQI-9 gap (AI − clinician) per session; higher = larger AI advantage. It **grows with noise** — clinician notes degrade faster than AI notes.

ROBUSTNESS

The effect survives every sensitivity check

Length-adjusted **+4.95**
AI notes are ~2× longer, but adjusting for length barely moves the gap (+5.08→+4.95); length itself is unrelated to quality (p=0.80).

Evaluator-outlier excluded **+4.86**
Dropping the most extreme scorers leaves the advantage essentially unchanged — not driven by a handful of raters.

Raw pre-calibration flags **same direction**
The safety result is already present in the raw judge output, before any false-positive calibration — not an artefact of tuning.

Also robust to specialty and acoustic condition; the omission advantage **widens** after length adjustment.

INSTRUMENT VALIDATION

Can you trust the detector? A pre-registered, blinded human-validation arm

10 external post-CCT clinicians · **565 adjudications** of 434 stratified pipeline flags (kept + FP-removed) · 131 double-rated · blinded to authorship, source, pipeline verdict and severity · binary genuine / not-genuine decision.

0.24
AC1 inter-clinician agreement — **no human gold standard exists**

64% vs 59%
judge–clinician ≈ clinician–clinician agreement — **within the inter-clinician envelope**

68%
latent-class genuine-error rate among flagged candidates (94% CI 48–83; Hui–Walter)

79%
human-confirmed kept-precision of the pipeline

PRE-SPECIFIED DECISION RULE — ALL THREE CRITERIA MET

A **Envelope parity** — judge agreement within the inter-clinician range. **Met.**

B **Non-differential across arms** — kept-precision AI 74% vs clin 81%; severity gap +0.06 vs −0.09 tiers. **Met.**

C **Concordance** — 70% on 77 clinician-consensus items. **Met.**

Why this makes the parent result conservative

A blinded, **non-differential** instrument — one that errs equally on AI and clinician notes — biases the AI-vs-clinician contrast **toward the null**. So the parent study's directional result is if anything **under-stated**, not inflated.

This is a **defensibility** claim about direction, not an accuracy claim about absolute error counts.

Conclusions

1. Across 5 languages and health systems, AI notes were simultaneously **higher-quality, more consistent and lower clinical-risk** than notes from the clinicians who author most routine documentation — from the same consultations.
2. The apparent **size** of the safety advantage is a function of detection sensitivity, which we measured under two regimes.
3. Human oversight is **least sensitive precisely on AI notes** — supporting automated, in-workflow verification over unaided human review.

Limitations

- Simulated actor-performed consultations (deliberate: gives ground truth).
- Unedited AI first draft assessed (a conservative floor).
- PDQI pooled ICC 0.24 — paired contrast robust, absolute values noisy.
- LLM-based error ground truth; validated against human performance, but no generalisable gold standard exists.
- Cross-site heterogeneity I^2=62.6%, driven by Germany's clinician baseline.
- Comparator is junior-to-middle-grade, not consultants. Validation arm English-only.

Competing interests

All authors are employed by **Heidi Health**, developer of the scribe and the automated reviewer.

Mitigations: authorship-blind pipeline; pre-specified SAP; result present in raw pre-calibration flags; corroborated by the independent human arm; calibration removed *more* AI flags than clinician flags.

Contact & manuscripts
henrybergman@heidihealth.com