

Beyond Dragon Medical Practice Edition: a browser-based generative AI platform for natural-speech radiology dictation — preliminary evaluation

Sanyam Katyal, MD, FRCR

Department of Radiology, University Hospitals of Derby and Burton NHS Foundation Trust, Derby, United Kingdom · Corresponding author: sanyam.katyal@nhs.net

POSTER CODE
rcrglobalai2026.178035c

STRUCTURED ABSTRACT

Purpose. To compare a new browser-based generative AI dictation platform against the incumbent Dragon Medical Practice Edition (DMPE) on word-error rate (WER), reporting time, cognitive workload and reader preference, in a within-subject crossover preliminary evaluation. **Methods.** Ten radiologists (trainee → consultant; subspecialties spanning abdominal, chest, neuro, MSK, cardiac and general imaging) each dictated 20 standardised cases (CT 8 / MRI 5 / US 2 / radiograph 3 / echocardiography 2) on both platforms in counterbalanced order, with a 14-day washout (n = 400 dictations). Primary endpoint: first-pass WER versus a fixed gold-standard report. Secondary: total time per case, post-correction WER, NASA-TLX, SUS and NPS. Paired analysis across (participant × case). **Results.** Mean WER 8.91 ± 7.05 % (DMPE) vs 2.20 ± 4.12 % (GenAI platform); ~75.3 % relative, p < 0.001 (paired Wilcoxon, n = 200 pairs), Cohen's d = 1.04. Mean total time 75.2 vs 26.0 s per case (~65.4 %); 93.0 % of the saving attributable to reduced correction effort. The GenAI platform outperformed DMPE in every participant and in all 20 cases. NASA-TLX frustration fell from 64 to 29; mean per-case satisfaction rose from 3.02 to 4.54 (1–5 scale). **Conclusion.** A browser-based generative AI dictation platform reduced WER by approximately three-quarters and total reporting time by approximately two-thirds compared with DMPE, with the largest gains in trainees and non-native speakers. A multi-centre prospective pilot, with adjudicated clinically-significant-error rate as the primary safety endpoint, is the appropriate next step.

01 Background

Speech recognition is the rate-limiting step of radiology reporting. Published per-report error rates range 4.8–89 % [1]; in one prospective audit 22 % of finalised reports contained errors, while 63 % of radiologists believed their personal rate was below 10 % [2]. The cost is paid in correction time, particularly for trainees and non-native speakers.

Generative AI offers a different architecture: real-time speech-to-text followed by an LLM that restores punctuation, repairs medical terminology and assembles structure automatically [3].

Hypothesis. A browser-based generative AI dictation platform reduces first-pass WER and total reporting time vs DMPE, with the magnitude of gain inversely related to prior DMPE familiarity.

03 Participants (n = 10)

ID	GRADE	SUBSPEC.	YRS	DMPE FAM
P01	Consultant	Abdominal	14	5/5
P02	Consultant	Chest	11	4/5
P03	Trainee	General	4	2/5
P04	Consultant	Neuro	18	5/5
P05	Fellow	MSK	6	3/5
P06	Trainee	General	2	2/5
P07	Consultant	General	22	5/5
P08	Consultant	Cardiac/US	9	4/5
P09	Consultant	Chest	16	5/5
P10	Locum Cons.	General	12	3/5

Median experience 11.5 yrs (IQR 6–16). 4/10 non-native English speakers. Recruited from the author's professional network, completed remotely. 10-minute familiarisation video before first GenAI dictation.

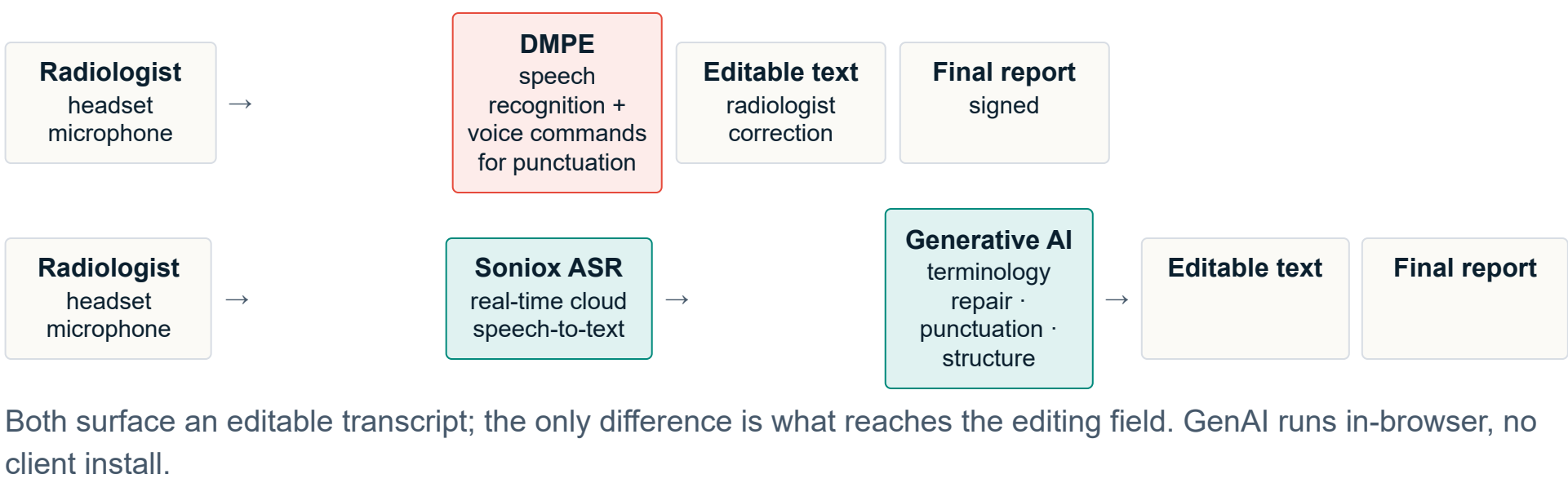
02 Methods

Design. Within-subject counterbalanced crossover (odd participants DMPE first; even, GenAI first). 14-day washout. Remote completion via the browser study app, same microphone for both arms.

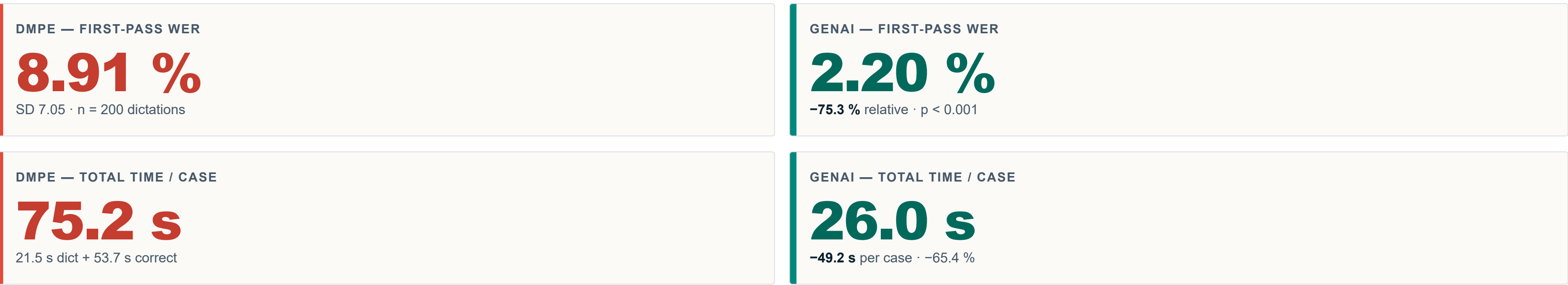
Cases. 20 standardised gold-standard reports curated by the author: CT 8 / MRI 5 / US 2 / radiograph 3 / echocardiography 2; 23–63 words. 17 pathology, 3 normal; one deliberately vocabulary-dense outlier.

Endpoints. Primary — first-pass WER (Levenshtein). Secondary — total/correction time, post-correction WER, satisfaction, NASA-TLX. Descriptive — edit-ops, SUS, NPS.

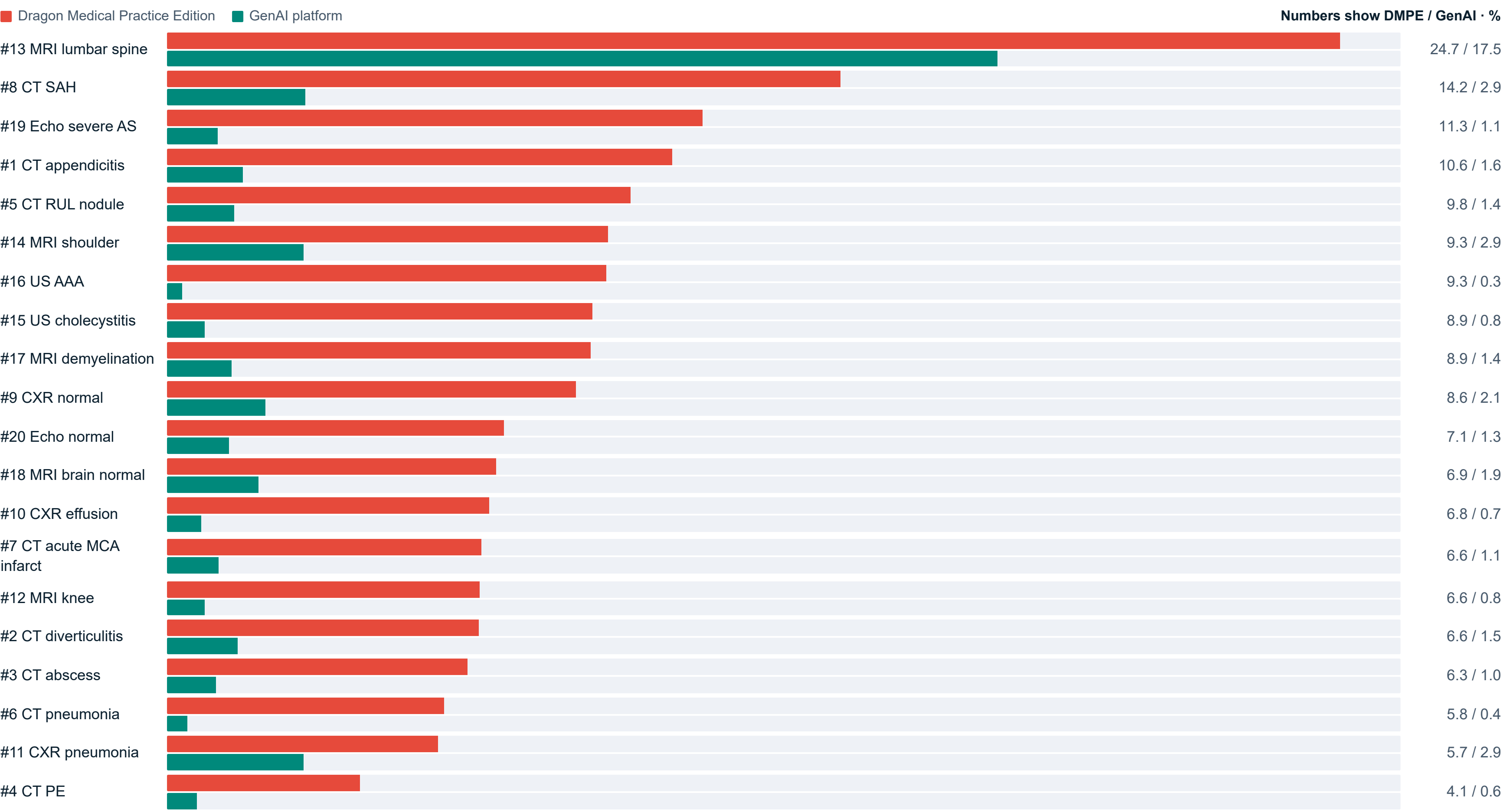
Analysis. Paired Wilcoxon signed-rank across n = 200 pairs; Cohen's d; Bonferroni across five secondary endpoints. All timings and edit events captured server-side.



04 Primary outcome & key effect sizes

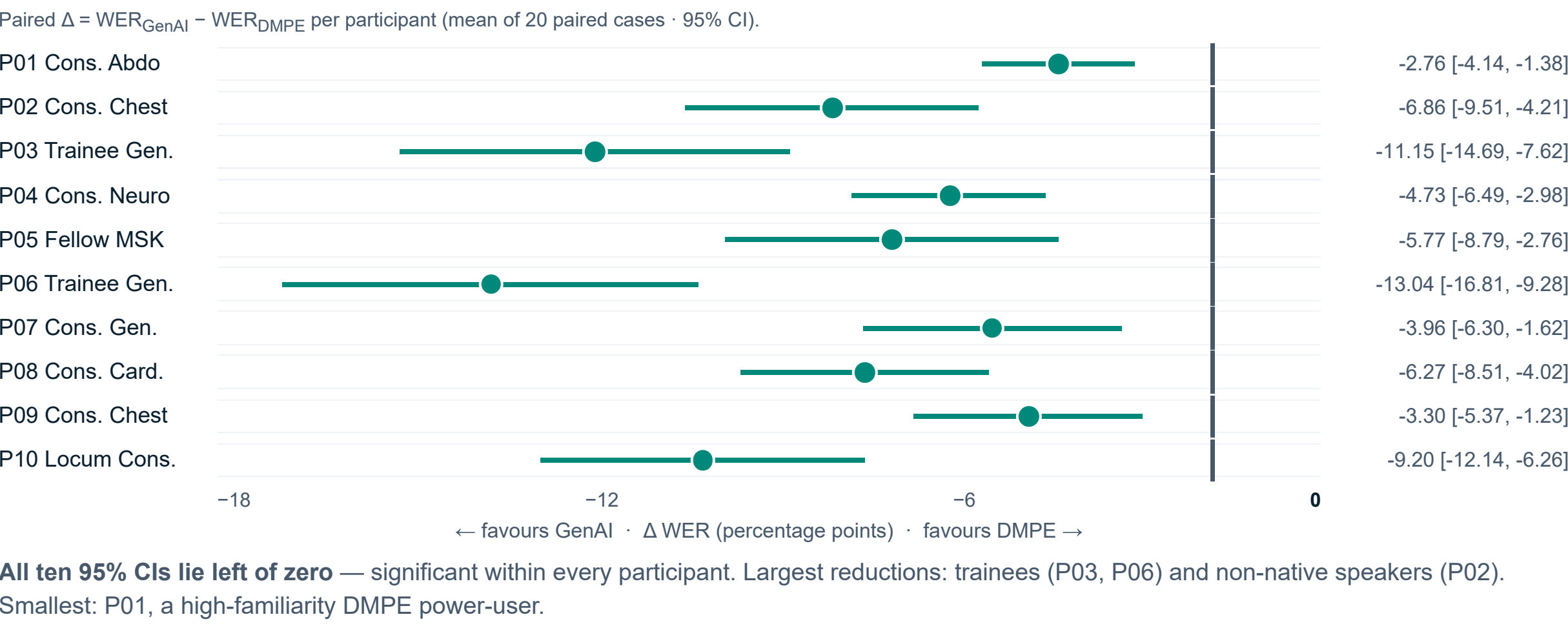


05 WER per case (mean across 10 radiologists)

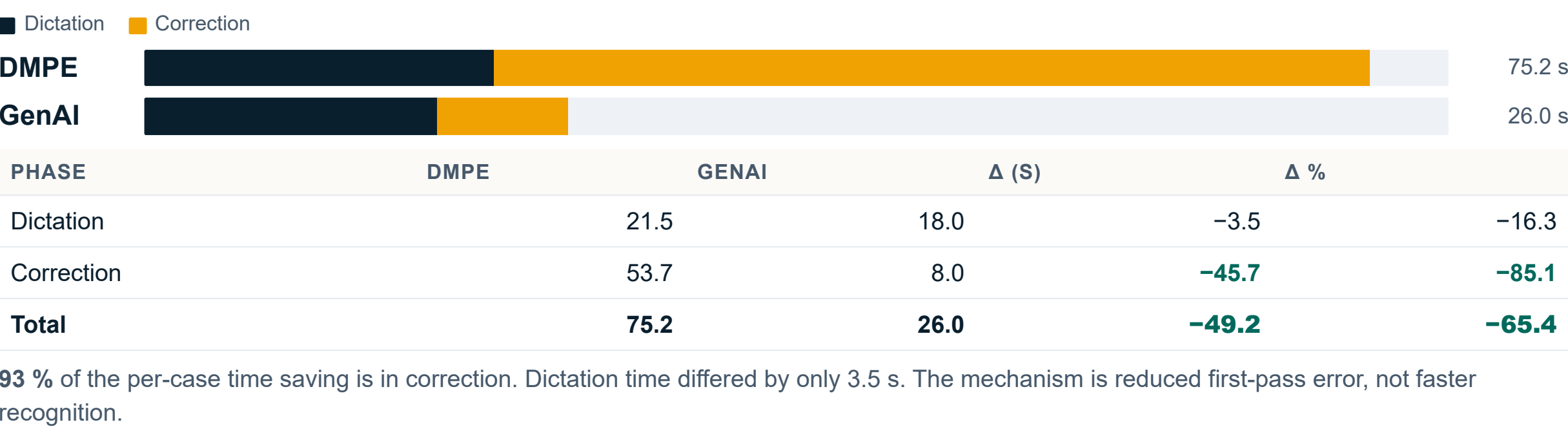


Outlier — Case 13 (MRI lumbar spine). Polysyllabic-dense report; both platforms struggle. GenAI retains a 7.2-pp advantage but 17.5 % absolute WER remains suboptimal — motivates the RadLex post-processor (§10).

06 Per-participant effect (forest plot)



07 Time decomposition



08 Statistical inference

ENDPOINT	Δ (GENAI - DMPE)	95 % CI	P	COHEN'S D
First-pass WER (%)	-6.71	-7.60, -5.81	< 0.001	1.04
Total time per case (s)	-49.1	-55.9, -42.4	< 0.001	1.01
Correction time (s)	-45.7	-52.3, -39.1	< 0.001	0.96
Post-correction WER (%)	-5.55	-6.48, -4.61	< 0.001	0.82
Per-case satisfaction (1–5)	+1.52	+1.31, +1.74	< 0.001	0.98
NASA-TLX frustration (0–100)	-35.4	-43.0, -27.8	< 0.01	2.89

Paired Wilcoxon, n = 200 (participant × case) pairs; TLX n = 10. Bonferroni-corrected. All effect sizes large; direction preserved in every participant.

09 Edit operations & subjective measures

Edit-operation breakdown (mean per dictation)			Subjective measures		
OP	DMPE	GENAI	MEASURE	DMPE	GENAI
Substitutions	1.98	0.53	Satisfaction (1–5)	3.02	4.54
Deletions	1.32	0.05	TLX mental	62	34
Insertions	0.32	0.24	TLX temporal	63	31
Total edits	3.62	0.82	TLX effort	66	39
Keystrokes	142.3	20.6	TLX frustration	64	29
Spoken-cmd ratio	17.0 %	—	SUS mean	3.2	4.2
			NPS	3	7

Cognitive load. DMPE requires every punctuation token to be spoken (*"period"*, *"new paragraph"*) — 17 % of spoken tokens. GenAI restores punctuation, capitalisation and structure automatically, allowing continuous natural-prose dictation; explains the simultaneous TLX reductions across all four subscales.

10 Discussion

- **Mechanism.** The cleanup pass repairs medical-terminology mis-recognitions, restores punctuation and assembles structure — yielding lower first-pass WER and offloading the punctuation cognitive burden (§09).
- **Equity.** DMPE quality scales with profile-training; GenAI is uniform across users because cleanup is downstream of recognition. Largest gains: trainees and non-native speakers. Earlier interventions (prepopulated templates) did not consistently reduce errors [4].
- **Shared failure mode.** Vocabulary-dense studies (Case 13) remain difficult for both. Next step: RadLex-anchored post-processing [5].
- **Architecture, not licensing.** Browser-based, no installation, no voice-profile training — productive from first session.
- **Comparator scope.** DMPE is widely deployed and the appropriate baseline here. Comparison against Nuance's cloud successor (DMO) is planned as a separate arm.

11 Limitations

- **Small preliminary cohort** (n = 10). External multi-centre validation required.
- **WER ≠ clinical safety.** Levenshtein WER weights a hallucinated finding equally to a misspelled word. Adjudicated clinically-significant-error rate is the appropriate primary endpoint for any safety study.
- **Standardised case set** cannot capture the full long tail of subspecialty vocabulary.
- **Single ASR / single LLM.** Generalisability across vendors and model versions not addressed.
- **Order effects.** Counterbalanced and washed-out, but residual learning cannot be fully excluded.
- **Comparator scope.** DMPE only; Nuance's cloud successor (DMO) not evaluated. Findings apply specifically to DMPE.

12 Conclusion & next step

A browser-based generative AI dictation platform produced substantially lower first-pass WER (2.20 % vs 8.91 %) and shorter reporting time (26.0 vs 75.2 s per case) than DMPE. The dominant mechanism was reduced correction effort. Direction of effect preserved across all 10 participants and all 20 cases, with the largest gains in trainees and non-native speakers.

NEXT STEP
Multi-centre prospective crossover pilot (n ≥ 30, ≥ 4 sites); three arms — DMPE, DMO, GenAI. Primary safety endpoint: adjudicated clinically-significant-error rate. Primary effectiveness endpoint: total per-case reporting time. RadLex post-processor in development.

SELECTED REFERENCES

1. Hammama I, Lepanto L, Puder T, Bellemare C, Ly M-S. Speech recognition in the radiology department: a systematic review. *Health Inf Manag* 2015;44(2):4–10. doi:10.1177/18333583150440201. PMID 26157081.
2. Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. *J Am Coll Radiol* 2008;5(12):1196–1199. doi:10.1016/j.jacr.2008.07.005. PMID 19027683.
3. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology* 2024;310(1):e232756. doi:10.1148/radiol.232756. PMID 36226883.
4. Hawkins CM, Hall S, Hardin J, Salisbury S, Towbin AJ. Prepopulated radiology report templates: a prospective analysis of error rate and turnaround time. *J Digit Imaging* 2012;25(4):504–511. doi:10.1007/s10278-012-9455-9. PMID 22270786.
5. Radiological Society of North America. RadLex. radlex.org. Accessed 2026.

DISCLOSURES & ETHICS

No financial relationships with Nuance or any ASR/LLM vendor. No external funding or institutional sponsorship. Independent investigator-initiated study; participants volunteered remotely after individual consent. HRA approval being sought for the planned prospective pilot.

ACKNOWLEDGMENTS

The author thanks the ten radiologists and trainees who voluntarily contributed their time, and the professional networks through which they were recruited.