

Symphony for Text Generation

Benchmarking Clinical Note Generation

Daniel Varab Victor Petréen Bach Hansen Matthew Newton Ashbjørn W Helge

Kevin Pelgrims Mathias Baltzersen* Adrian Young-San Roessler Vanessa Klungtvedt

Maximilian Brand Lasse Krogsbøll Henrik Cullen Lars Maaløe

Abstract

Ambient documentation systems are rapidly gaining adoption, yet their impact on clinical note quality remains poorly characterized. We introduce MedConv, a multilingual dataset of 300 clinical encounters in English, Danish, and German, and use it alongside the Ambient Clinical Intelligence benchmark (ACI-BENCH) to compare Corti, a clinical AI platform, with two leading, accessible ambient scribe software applications built on general-purpose AI. We present a controlled clinical evaluation framework that combines entailment metrics with LLM-judged pairwise comparisons across eight dimensions adopted from PDSQI-9. Results show that Corti’s API-based text-generation infrastructure is on par with or outperforms leading commercial scribes. We further show that Corti’s configurable API provides the flexibility necessary to fine-tune quality dimensions for specific documentation use cases. We present the evaluation methodology and release a dataset to support future reproducible comparison of ambient documentation systems.

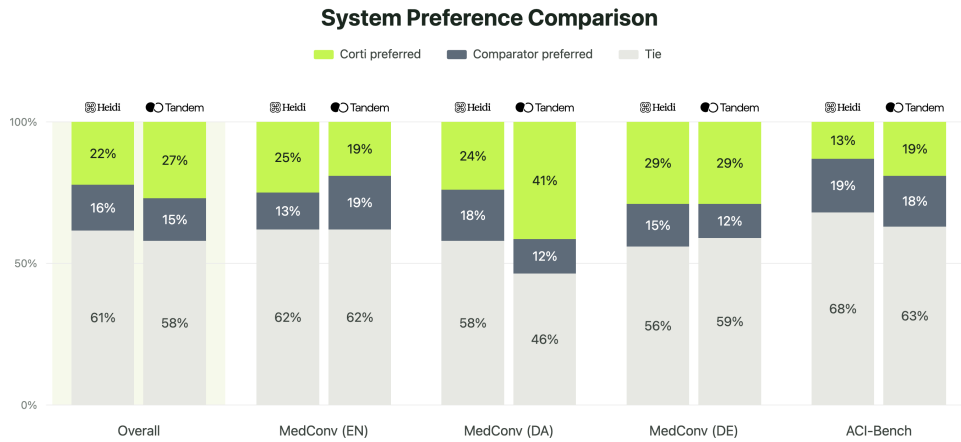


Figure 1: Pairwise preference outcomes on SOAP notes, shown for each dataset and overall pooled. For each Corti–comparator pair, bars report the percentages of judgments preferring Corti, preferring the comparison system, or resulting in a tie, aggregated over the eight adapted PDSQI dimensions. Contradictory judgments are counted as ties. The pooled result comprises 3,296 dimension-level judgments per compared system (412 encounters $\times$ 8 dimensions).

*Correspondence to mbn@corti.ai, Corti. To get access to Symphony for Text Generation, visit Corti Docs.

1 Introduction

Ambient documentation systems are increasingly used to reduce clinical documentation burden and recent studies report improvements in clinician experience, well-being, and documentation efficiency across a range of deployment settings [van Linschoten et al., 2026, Olson et al., 2025, Pearlman et al., 2025, Lukac et al., 2025, Afshar et al., 2025]. As the adoption and workflow benefits improve, the quality of the clinical documentation remains largely under-reported.

Clinical note quality is difficult to compare systematically across ambient documentation systems: existing cross-vendor studies vary in scale and in how tightly they control the inputs and configurations each system receives [Anderson et al., 2025, Draper et al., 2025, Fox et al., 2026a]. Omissions as a dominant, critical error type with potentially severe patient impact, are especially hard to measure, since evaluators that inspect only what a note contains can miss what it left out [Draper et al., 2025, Schumacher et al., 2025, Fox et al., 2026b].

Surface-level measures, such as BLEU [Papineni et al., 2002], METEOR [Banerjee and Lavie, 2005] and ROUGE [Lin, 2004], are poorly suited to clinical text (§E.1), and recent approaches, therefore, combine semantic or factual assessment, clinician review, structured quality instruments, and LLM-based evaluation to resolve their shortcomings [Wang et al., 2025, Croxford et al., 2025a,b].

As a result, teams building ambient scribes (or health systems trying to select one) are forced to rely on qualitative surveys and anecdotal evidence. This creates the risk of conflating the performance of underlying AI platform with factors that are extraneous to it, such as workflow integration, user interface design, or the preferences of a lead clinician on a pilot team.

In this report, we present a methodology and a public dataset for evaluating ambient documentation performance. We demonstrate how a purpose-built AI platform matches or outperforms commercial ambient scribe software applications built on general-purpose AI platforms while enabling simple, safe tailoring to clinician preferences. The goal is to provide anyone building AI solutions for clinical documentation a way to easily tailor documentation towards specific workflow needs and quantify the resulting performance.

Our contributions are as follows.

- **A multi-dimensional evaluation protocol for clinical note generation.** We score each note on three automatic entailment measures: groundedness (*is it supported by the consultation?*), completeness (*does it include what the reference note includes?*), and conciseness (*does it include grounded content that’s omitted from the reference?*). We also use LLM judges to compare notes from two systems side-by-side using validated clinical criteria adopted from PDSQI-9, with clinicians further validating a subset of the judgments for correctness.
- **A multilingual clinical documentation dataset.** We introduce MedConv, a dataset comprising of 300 diverse synthetic clinical scenarios with matched transcript–note pairs in English, Danish, and German. Designed to support controlled evaluation across clinical domains and encounter settings.
- **A controlled multilingual cross-system benchmark.** We evaluate two widely adopted ambient documentation applications built on general purpose AI platforms, Heidi and Tandem Health, and the vertical AI platform Symphony by Corti on the MedConv and ACI-BENCH datasets.
- **An empirical analysis of template configurability.** We show that Corti’s configurable AI platform enables small, targeted changes in prompt design to meaningfully shift quantitative performance, allowing clinicians to tailor output quality along the dimensions most relevant to their use case.

2 Methodology

2.1 Experimental Design

We design a paired benchmark to compare clinical note quality and evaluate three candidate systems. To focus the comparison on note generation, we consider only text transcripts as input, excluding potential speech recognition errors. Systems generate notes from identical pre-transcribed inputs under a harmonised four-section SOAP template, using default generation settings for each system. All evaluation measures are computed at encounter level. SOAP is the only note type evaluated, being the format most consistently supported across the three systems and three markets. Notes are assessed through complementary evaluation, measuring both content fidelity and fine-grained documentation quality. We consider both reference-based, and reference-free metrics to measure coverage of notes, as well as paired comparisons to assess the relative quality of notes generated from the same encounter across domain specific quality dimensions.

2.2 Evaluation Datasets

To compare systems we use two datasets: the publicly available ACI-BENCH corpus [Yim et al., 2023], and MedConv, a dataset created by Corti that extends coverage to additional medical domains, consultation settings, and languages (see Table 1).

Table 1: Dataset sizes, mean transcript length, and estimated durations at 150 words-per-minute.

Dataset	N	#words (min–max)	Avg. encounter duration (min)
ACI-BENCH (EN)	112	1271 (710–2541)	8.5
MedConv (DA)	100	1655 (923–2585)	11.0
MedConv (DE)	100	1822 (904–2739)	12.1
MedConv (EN)	100	2031 (930–3073)	13.5

ACI-BENCH The Ambient Clinical Intelligence Benchmark (ACI-BENCH) corpus, published by Microsoft Health AI, Nuance Communications, and University of Washington, comprises of 207 cases across three subsets representing different modes of clinical note generation from doctor-patient conversations[Yim et al., 2023]. We use the *aci* subset, since it most closely matches the ambient-scribe scenario evaluated in this benchmark. While the dataset does not come with explicit medical domain labels, LLM-assisted analysis points to diverse coverage across 10 domains, with musculoskeletal/orthopedic dominating and cardiovascular, respiratory, gastrointestinal, and hepatology well represented. Note structures are diverse; the original authors grouped them into four standardized meta-divisions, Subjective, Objective-exam, Objective-results, and Assessment-plan, which align closely with the SOAP structure adopted for this work.

MedConv The Corti MedConv corpus comprises of 100 transcript-notes pairs per language (English, Danish, German; 300 cases total) spanning 15+ medical specialities. 100 English notes are generated by seeding a synthetic pipeline with clinician authored stories and clinical clues. Each clinical note is validated and corrected by the same clinicians. Transcripts are generated synthetically using another synthetic pipeline which conditions on the validated clinical note, constructing a noisy multi-speaker consultation transcript. The process is repeated for each target languages. Encounter settings are mixed: outpatient (32), inpatient (24), admission (19), general practice (11), pre-hospital (7), and investigation (7). Reference note structures are diverse and tailored to individual cases and domains, and are not generated according to Corti’s templates.¹

¹The MedConv dataset will be released publicly alongside a paper detailing its construction and validation. Upon release, the dataset will be available at <https://huggingface.co/datasets/corti/med-conv>.

2.3 Systems Under Review

Heidi and Tandem Health Heidi² and Tandem Health³ are stand-alone clinician-facing ambient documentation software applications, supporting template-configured clinical note generation. We select these clinical systems, rather than general-purpose model APIs, such as OpenAI’s GPT models or Google’s Gemini model family, as they represent some of the fastest growing, industry-leading software application integrations of such models in healthcare contexts⁴. The challenges identified in these well-resourced software applications are likely to be even greater for developers of software applications with fewer resources.

Corti Corti’s Symphony offers a clinical AI platform that includes purpose-built capabilities for speech-to-text [Nix et al., 2026], text generation, medical coding [Edin et al., 2026], agents, and general model inference⁵. Its text-generation capabilities provide an API-based alternative to using general-purpose LLMs directly, supporting transcript- and/or fact-based document synthesis with configurable, section-level templates and structured outputs. For this evaluation, Corti was accessed through Symphony’s text-generation APIs.⁶

For a full reference of the configurations used for each system, see §C in the appendix.

Note Generation Notes are generated in September 2026 using each system’s available interface: Corti via the Text Generation API⁷, and Heidi and Tandem Health manually through their end-user web interface - providing transcripts as consultation context and exporting generated notes as plain text. For Corti, we sample five notes per encounter across all four datasets through the API. Because Heidi and Tandem Health require manual generation through web interfaces, it was intractable to run large-scale comparisons. As a result, five notes per encounter were sampled for each system on ACI-BENCH, and one per encounter on the MedConv datasets. This allows measuring variation across systems while recognising that variation may not generalise to MedConv. Within each system, inputs, templates, and configurations, are fixed across replicates, such that between-run variation reflects only stochastic nature of the generation pipelines.

Configuration and template harmonization We consider one canonical SOAP note type per system and geography, harmonizing headings as far as each template editor allowed. Changes are limited to section reordering, heading renames, and splitting or merging sections; we do not rewrite vendor prompt text (Table 5, Table 6). We record output-affecting settings and fix them to vendor defaults, with Corti in its default production configuration (Table 4). Market configurations were UK English, German, and Danish; the US-origin ACI-BENCH cases are evaluated under the UK configuration.

2.4 Evaluation Measures

Metrics Surface-level measures such as BLEU, ROUGE, and METEOR assume that overlap in wording implies proximity in semantics. Clinical documentation violates in both directions: equivalent findings are routinely expressed with disjoint vocabulary, and a negation or reversed causal relation can render a note false while leaving nearly all wording intact. Because errors of the latter kind are lexically conservative, an n -gram scorer will rank a clinically incorrect note above an acceptable paraphrase, which makes these measures least reliable on precisely the failure modes that carry the greatest clinical risk.

For this reason, similar to Hansen et al. [2025], we adopt a model-as-a-judge approach when evaluating the quality of the generated notes. An LLM is used to assess the quality of generated outputs in place of traditional automatic metrics or human annotators [Gao et al., 2024, Li et al., 2024, Yehudai et al., 2025].

²<https://www.heidihealth.com/>

³<https://www.tandemhealth.ai/>

⁴Other commercial scribes were not considered due to access limitations or limited language coverage.

⁵<https://docs.corti.ai>

⁶<https://docs.corti.ai/textgen>

⁷See §B in the appendix for a detailed description of the full pipeline

Given the multi-faceted complexity of evaluating clinical note generation, we propose a multi-dimensional evaluation framework built on two complementary judgments. The first is reference-based: does the generated note capture the clinically relevant content of the reference note without adding content beyond it, and is everything it states supported by the transcript? We operationalize this as three entailment-based metrics adapted from Hansen et al. [2025], in which an LLM judge checks whether each statement in one text is supported by another: reference statements against the generated note (completeness), generated statements against the reference (conciseness), and generated statements against the transcript (groundedness).

Entailment scores capture content fidelity but not the dimensions clinicians weigh when judging nuanced note quality. PDSQI-9, introduced by Croxford et al. [2025a,b], is a clinician-validated instrument to measure the quality of clinical notes in a set of fixed dimensions (a.o. organization, synthesis, succinctness, and stigmatizing language). We extend the instrument to a preference-based measure and compare notes pairwise across eight of the instruments dimensions. To reduce position bias, we evaluated each note pair in both orders. We assign system preference only when both judgments agree, and otherwise record a tie (§E.2). We validate the extension against blinded clinician preferences on a targeted case subset (§G) and include document judge-model sensitivity of the measure in (§E.3). We provides a preference measure that allows direct evaluation of note-generation pipelines and, as we show in §3.3, can be used to monitor improvements and regressions related to scribe or template configuration changes.

For detailed definitions we refer to the appendix in §E.1 for the entailment-based metrics and §E.2 for the pairwise-adapted PDSQI-9 dimensions.

Latency Runtime was recorded for each generated note: API-side for Corti, from request submission to receipt of the complete generated document, and browser-side for comparators, from submission to the fully rendered note. Measurements were taken sequentially from Copenhagen in early September 2026.

Statistical analysis Pairwise preference scores are reported with 95% confidence intervals from a non-parametric cluster bootstrap over encounters (§E.2); intervals including 50 indicate no detectable preference. We report all eight dimensions without multiplicity correction, so dimension-level results are descriptive rather than independent hypothesis tests. The two families’ uncertainty estimates are not interchangeable: one captures generation variability without encounter sampling, the other the reverse.

Choice of judge Pairwise preferences depend on the judge model and are sensitive on the chosen models biases. We assess sensitivity by computing comparisons with OpenAI’s GPT-5.4 and Anthropic’s Opus 4.6 (§E.3). We observe similar preference distributions, however, with Opus 4.6 generally assigning more ties. We selected GPT-5.4 as the primary judge for all model-based metrics as it yielded fewer indeterminate comparisons.

2.5 Clinical Validation of the Pairwise Evaluation

We further employ human evaluation to validate preference measures. We task three in-house clinicians, with diverse backgrounds (German, Norwegian, Danish) across multiple specialties, to validate the LLM verdicts on each of the PDSQI-dimensions. 10 notes from the ACI-BENCH runs are sampled and a blinded pairwise comparison of the notes generated by two systems are presented in a simple web-based user interface to each clinician separately. See Figure 6 and 7 in the appendix for the interfaces presented to the clinicians. For each of the sample cases, the user interface presented a side-by-side view of the two notes, randomly flipped to avoid position bias, and the reasoning traces of both position-swapped judgments for the dimensions.

Clinicians endorse the judge’s verdict in 207 of 240 assessments (86.3%, Gwet’s AC1 = 0.694), and a majority of the three clinicians are in agreement in 75 of 80 case-dimension comparisons, supporting the validity of the pairwise judge as a scalable proxy for clinician preference (§G). Agreement was highest for accuracy, thoroughness, and usefulness and lowest for synthesis (AC1 = 0.080).

Table 2: Content related metrics across datasets. Highest scores within each dataset are shown in bold. $\pm$ denotes standard deviation over means across multiple runs; – marks single-run entries.

Dataset	System	Groundedness $\uparrow$	Completeness $\uparrow$	Conciseness $\uparrow$
ACI-Bench (EN)	Corti	96.3% $\pm$ 0.3%	77.3% $\pm$ 0.2%	74.9% $\pm$ 0.2%
	Heidi	94.9% $\pm$ 0.4%	75.9% $\pm$ 0.3%	75.7% $\pm$ 0.2%
	Tandem Health	94.7% $\pm$ 0.3%	73.1% $\pm$ 0.4%	74.0% $\pm$ 0.1%
MedConv (EN)	Corti	97.8% $\pm$ 0.1%	73.0% $\pm$ 0.1%	81.4% $\pm$ 0.3%
	Heidi	95.4% $\pm$ –	64.6% $\pm$ –	83.0% $\pm$ –
	Tandem Health	96.7% $\pm$ –	62.9% $\pm$ –	81.3% $\pm$ –
MedConv (DA)	Corti	97.1% $\pm$ 0.3%	71.9% $\pm$ 0.4%	85.7% $\pm$ 0.2%
	Heidi	94.8% $\pm$ –	64.1% $\pm$ –	85.6% $\pm$ –
	Tandem Health	89.7% $\pm$ –	61.8% $\pm$ –	78.8% $\pm$ –
MedConv (DE)	Corti	98.1% $\pm$ 0.2%	74.5% $\pm$ 0.2%	82.6% $\pm$ 0.2%
	Heidi	93.9% $\pm$ –	65.1% $\pm$ –	84.1% $\pm$ –
	Tandem Health	96.7% $\pm$ –	65.3% $\pm$ –	82.9% $\pm$ –

3 Results

3.1 Overall performance evaluation of the three systems

Corti produces the most complete notes on all four datasets, while the three systems score similarly on groundedness and conciseness (Table 2). Here, *completeness* measures how much of the clinically relevant information in the reference note is captured, *groundedness* measures whether statements in the generated note are supported by the source transcript, and *conciseness* measures whether the generated note avoids bloat by outputting grounded content absent from the reference.

Completeness is the lowest-scoring dimension for all systems (62–77%) and the clearest source of differentiation. Corti’s advantage ranges from 1 percentage point on ACI-BENCH to 8–10 points on MedConv, suggesting that the systems diverge more on the more challenging MedConv cases. By contrast, groundedness is consistently high (90–98%), and both groundedness and conciseness are generally within two points across systems; Tandem Health on MedConv Danish is the main exception scoring lower. Run-to-run standard deviations are at most 0.4 points across five repetitions (for all systems on ACI-BENCH and Corti on MedConv) suggesting consistent outputs. An exploratory breakdown of the flagged statements by failure type and severity is provided in §I of the appendix.

Corti has the lowest mean latency across all datasets, at 5.9 to 9.7 seconds per note, while Heidi took 2.0 to 2.7 times as long and Tandem Health 1.9 to 2.2 times as long (Table 3).

Table 3: Speedup ratios and dataset statistics relative to Corti.

Dataset	Transcript chars	Corti	Heidi		Tandem Health	
		avg (s)	avg (s)	vs Corti	avg (s)	vs Corti
ACI-BENCH (EN)	6101 (3474–12097)	5.86	15.59	2.7 $\times$	11.33	1.9 $\times$
MedConv (DA)	7591 (4243–12115)	9.68	19.68	2.0 $\times$	21.41	2.2 $\times$
MedConv (DE)	9551 (5269–15258)	9.07	19.22	2.1 $\times$	18.97	2.1 $\times$
MedConv (EN)	8960 (4973–13949)	7.49	17.60	2.3 $\times$	14.18	1.9 $\times$

3.2 Granular evaluation on clinical dimensions

A pairwise preference evaluation provides a more granular view of where the systems differ clinically. Results are displayed in Figure 2. Rather than reducing note quality to aggregate metrics, it compares outputs across rich dimensions such as accuracy, thoroughness, usefulness, organization, comprehensibility, synthesis, and succinctness. This reveals trade-offs that are not apparent from the entailment-focused evaluation alone: systems with similar overall groundedness or conciseness can still differ substantially in how complete, useful, well-structured, or appropriately synthesized their notes are.

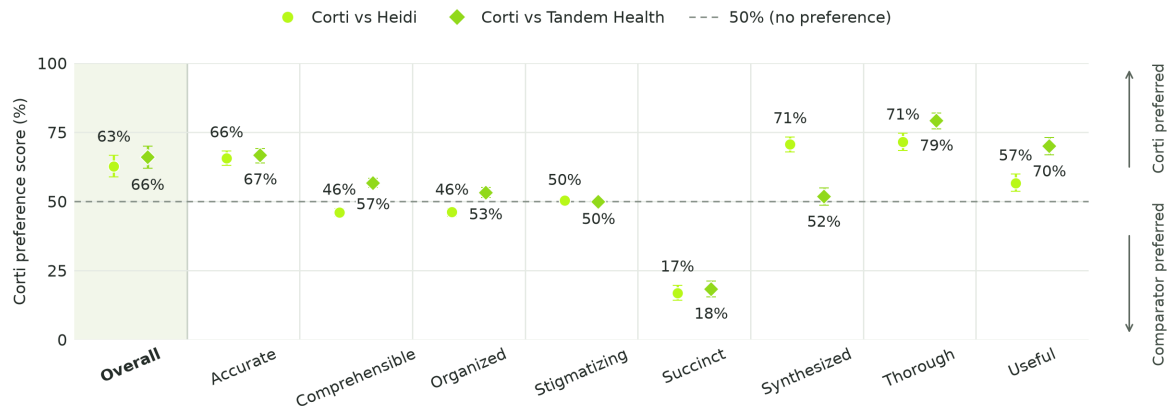


Figure 2: PDSQI-9 preference scores on SOAP, comparing Corti with each evaluated system.

Pooled across datasets, the LLM judge prefers Corti’s notes over both Heidi’s and Tandem Health’s (Figure 1). Corti is preferred 22% of the evaluations against Heidi at 16% and 27% against Tandem Health at 15%. The only case where an ambient scribe application is preferred against Symphony is on ACI-BENCH where Heidi is preferred 19% of the evaluations against Corti at 13%. The high proportion of ties indicates well-calibrated and high-performing systems overall (46–68%).

Across all dimensions, Corti wins overall 63% and 66% of the time against Heidi and Tandem Health respectively.⁸ Against both comparators, Corti is preferred on **accuracy**, **thoroughness**, and **usefulness**, and the comparator is preferred on **succinctness** (Figure 2). Corti’s largest advantage is on thoroughness (71.5 against Heidi, 79.1 against Tandem Health) and its largest deficit is on succinctness (16.9 and 18.2). The comparator-dependent dimensions split as follows: against Heidi, Corti is preferred on synthesis (70.6) but slightly less preferred on comprehensibility and organization (both 46); against Tandem Health, Corti is preferred on comprehensibility and organization and roughly neutral on synthesis (51.7).

3.3 Targeted intervention to improve on succinctness

Different documentation contexts demand different levels of granularity. To test a targeted intervention on improving the succinctness dimension for the Corti API, we test a template optimized for General Practice which equates some simple changes to the prompts in the template. We add the following general-practice instruction:

This is a general-practice visit note. Keep it brief and deliberately selective: document only the clinically salient, decision-relevant information and leave out the rest. Synthesize the source into short, covering statements to achieve a correct but easily readable note.

Furthermore, we change the writing-style configuration from a comprehensive, fluent style using sentences of up to 20 words to a terse, detailed style using telegraphic statements of up to 15 words, and added compression instructions into the subjective section:

Aim for 3–4 short sentences. The default is to OMIT

We repeat the pairwise evaluation against the comparators. See Table 7 for more details on templates.

Figure 3 displays the preference scores under this change. The succinctness preference score increases from 16.9% to 65.3% against Heidi. Against Tandem Health, it increases from 18.2% to 59.7%, making the Corti notes the preferred option also on the succinctness dimension.

As a trade-off, thoroughness score decreases but stays favourable for Corti above 50%. Other dimensions do not change uniformly. Against Heidi, organization decreases from 46.1% to 35.8% and comprehensibility from 46.0% to 40.5%. Against Tandem Health, synthesis changes from 51.7% to 48.1%.

⁸Refer to §E.2 in the appendix for the definition of the *overall* win-rate.

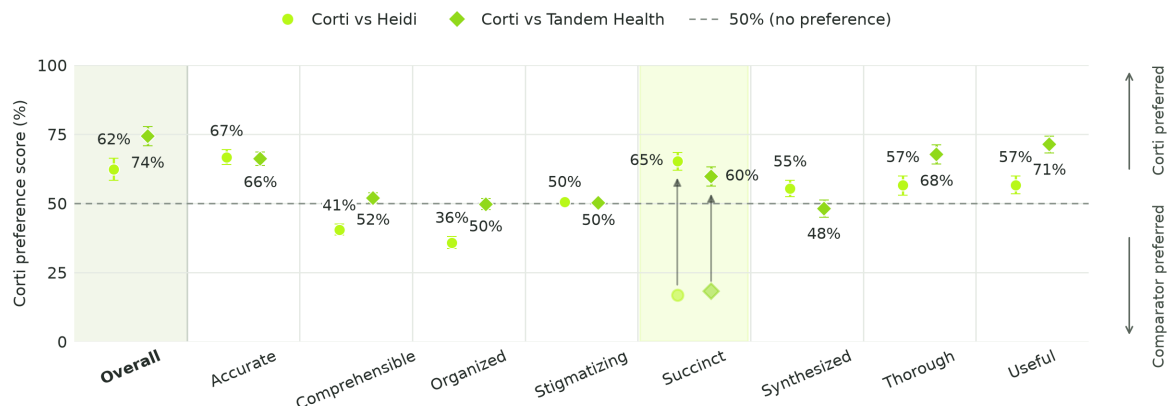


Figure 3: PDSQI-9 preference scores after adding general-practice instructions, comparing Corti with each evaluated system. Each dot represents a comparison between Corti and one other system.

By modifying only two narrowly scoped instruction fields while leaving the upstream base section unchanged, the intervention illustrates the practical value of section-level over monolithic template definitions

4 Conclusion

To address the growing need for multifaceted evaluations of ambient clinical notes in light of patient safety concerns, we present a controlled, multilingual framework across groundedness, completeness, conciseness, and multidimensional note quality. The evaluated systems differ in completeness and pairwise quality preferences. Corti achieves the highest completeness and groundedness scores across all four datasets, at the fastest generation runtime, and is preferred over both comparators on average across the clinical dimensions of the PDSQI framework.

The relative importance of completeness and succinctness depends on the clinical context: in complex or specialty encounters, where omissions may carry greater risk, completeness may be prioritized, whereas in high-volume, lower-complexity settings clinicians may favour brevity to reduce review burden. A small, surgical template intervention shifted Corti over comparators in succinctness while preserving an above-neutral preference for thoroughness, demonstrating that documentation behaviour can be steered through API-configurations by integrators, e.g., health provider or software integrator. This highlights that ambient documentation should be evaluated not only as a fixed product, but as a configurable systems whose optimal quality profile depends on the clinical setting and workflow.

More broadly, with this study, we highlights the importance of reproducible, programmatic evaluation. By releasing MedConv and presenting our evaluation framework, we aim to make comparisons of ambient documentation systems more transparent, controlled, and repeatable, to support health systems and scribe builders in evaluating both note quality and the degree of control they have over it.

5 Acknowledgements

We thank Jonas Lyngsø and Fabio Voermans for the design and implementation of this text generation approach and powerful API, Chituru Chinwah and the Corti AI Platform team for assuring quality and enabling a highly performant infrastructure, Robert James for making the MedConv dataset a reality. Thank you to the rest of Corti for their valuable contributions, insightful discussions, and support throughout the development and evaluation of this work.

References

- R. C. A. van Linschoten, C. M. van Loon, L. Joanknecht, et al. Ambient scribe in general practice: a multi-perspective before-after longitudinal mixed-methods study. *npj Digital Medicine*, 9:299, 2026. doi: 10.1038/s41746-026-02454-3.
- Kristine D. Olson, Daniella Meeker, Matt Troup, Timothy D. Barker, Vinh H. Nguyen, Jennifer B. Manders, Cheryl D. Stults, Veena G. Jones, Sachin D. Shah, Tina Shah, and Lee H. Schwamm. Use of ambient AI scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10):e2534976, 2025. doi: 10.1001/jamanetworkopen.2025.34976. URL <https://doi.org/10.1001/jamanetworkopen.2025.34976>.
- Kevin Pearlman, Wen Wan, Sachin Shah, and Neda Laiteerapong. Use of an AI scribe and electronic health record efficiency. *JAMA Network Open*, 8(10):e2537000, 2025. doi: 10.1001/jamanetworkopen.2025.37000. URL <https://doi.org/10.1001/jamanetworkopen.2025.37000>.
- Paul J. Lukac, William Turner, Sitaram Vangala, Aaron T. Chin, Joshua Khalili, Ya-Chen Tina Shih, Catherine Sarkisian, Eric M. Cheng, and John N. Mafi. Ambient AI scribes in clinical practice: a randomized trial. *NEJM AI*, 2(12), 2025. doi: 10.1056/AIoa2501000. URL <https://doi.org/10.1056/AIoa2501000>.
- Majid Afshar, Mary Ryan Baumann, Felice Resnik, Josie Hintzke, Anne Gravel Sullivan, Graham Wills, Kayla Lemmon, Jason Dambach, Leigh Ann Mrotek, Mariah Quinn, Kirsten Abramson, Peter Kleinschmidt, Thomas B. Brazelton, Margaret A. Leaf, Heidi Twedt, David Kunstman, Brian Patterson, Frank Liao, Stacy Rasmussen, Elizabeth S. Burnside, Cherodeep Goswami, and Joel Gordon. A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. *NEJM AI*, 2(12), 2025. doi: 10.1056/AIoa2500945. URL <https://doi.org/10.1056/AIoa2500945>.
- 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. URL <https://doi.org/10.1016/j.mcpdig.2025.100292>.
- Thomas C. Draper, Timothy Cox, Kathryn Lamb-Riddell, Luigi Andrea Moretti, John McCormick, Stephen Trowell, Janice Kiely, and Richard Luxton. Clinical AI scribes in primary care: accuracy, error severity and implications for clinical practice. *BMJ Digital Health & AI*, 1(1): e000092, 2025. doi: 10.1136/bmjdhai-2025-000092. URL <https://doi.org/10.1136/bmjdhai-2025-000092>.
- 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. arXiv preprint arXiv:2608.31017, August 2026a. URL <https://arxiv.org/abs/2608.31017v1>. Version 1, 31 August 2026.
- Elliot Schumacher, Daniel Rosenthal, Dhruv Naik, Varun Nair, Luladay Price, Geoffrey Tso, and Anitha Kannan. MED-OMIT: Extrinsically-focused evaluation metric for omissions in medical summarization. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pages 897–922. PMLR, 2025. URL <https://proceedings.mlr.press/v259/schumacher25a.html>.
- Sebastian Fox, Luke Markham, Ryan Lail, and Michael Karotsieris. LLM judges verify presence, not absence: Omission blindness in AI clinical notes and what recovers it. arXiv preprint arXiv:2608.31016, August 2026b. URL <https://arxiv.org/abs/2608.31016v1>. Version 1, 31 August 2026.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Haoyuan Wang, Rui Yang, Mahmoud Alwakeel, Ankit Kayastha, Anand Chowdhury, Joshua M. Biro, Anthony D. Sorrentino, Jessica L. Handley, Sarah Hantzmon, Sophia Bessias, Nicoleta J. Economou-Zavlanos, Armando Bedoya, Monica Agrawal, Raj M. Ratwani, Eric G. Poon, Michael J. Pencina, Kathryn I. Pollak, and Chuan Hong. An evaluation framework for ambient digital scribing tools in clinical applications. *npj Digital Medicine*, 8(1):358, 2025. doi: 10.1038/s41746-025-01622-1. URL <https://www.nature.com/articles/s41746-025-01622-1>.
- Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris Ebby, Jillian Gorski, Matthew Kalscheur, Samy Khalil, Marie Pisani, Tyler Rubeor, Peter Stetson, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Development and validation of the provider documentation summarization quality instrument for large language models. *Journal of the American Medical Informatics Association*, 32(6):1050–1060, 2025a. doi: 10.1093/jamia/ocaf068. URL <https://doi.org/10.1093/jamia/ocaf068>.
- Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, Matthew M. Churpek, Anoop Mayampurath, Frank Liao, Cherodeep Goswami, Karen K. Wong, Brian W. Patterson, and Majid Afshar. Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine*, 8(1):640, 2025b. doi: 10.1038/s41746-025-02005-2. URL <https://www.nature.com/articles/s41746-025-02005-2>.
- Wen-wai Yim, Yajuan 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.
- Arne Nix, Robert James, Lasse Borgholt, Anna B. Ekner, Lana Krumm, Julius Severin, Dan Engel, Lars Maaløe, and Jakob Havtorn. Symphony for speech-to-text: Supporting real-time medical voice interfaces, 2026. URL <https://arxiv.org/abs/2605.16545>.
- Joakim Edin, Andreas Motzfeldt, Simon Flachs, and Lars Maaløe. Symphony for medical coding: A next-generation agentic system for scalable and explainable medical coding, 2026. URL <https://arxiv.org/abs/2603.29709>.
- Victor Petrén Bach Hansen, Lasse Krogsbøll, Jonas Lyngsø, Mathias Baltzersen, Andreas Motzfeldt, Kevin Pelgrims, and Lars Maaløe. Factsr: A safer method for producing high quality healthcare documentation. *arXiv preprint arXiv:2505.10360*, 2025.
- Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. Llm-based nlg evaluation: Current status and challenges, 2024. URL <https://arxiv.org/abs/2402.01383>.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llm-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL <https://arxiv.org/abs/2412.05579>.
- Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of llm-based agents, 2025. URL <https://arxiv.org/abs/2503.16416>.
- Anand Chowdhury, Michele Casey, Jonathan Wilson, Kathryn I. Pollak, Benjamin A. Goldstein, Armando Bedoya, and Eric G. Poon. Comparing ambient scribes: a randomized crossover clinical trial addressing ambient scribe technologies’ impact on physician burnout. *Journal of the American Medical Informatics Association*, 33(5):990–999, 2026. doi: 10.1093/jamia/ocag018. URL <https://doi.org/10.1093/jamia/ocag018>.
- Lisa S. Rotenstein, A. Jay Holmgren, Robert Thombly, Aditi Sriram, Reema H. Dbouk, Melissa Jost, Debbie Aizenberg, Scott MacDonald, Naga Kanaparthi, Brian Williams, Allen Hsiao, Lee Schwamm, Sara Murray, Maria Byron, Jacqueline G. You, Amanda J. Centi, Christine Iannaccone, Michelle Frits, Adam B. Landman, Karandeep Singh, Ming Tai-Seale, Jie Cao, Katharine Lawrence, Devin Mann, Christopher Holland, Bryan Blanchette, Jesse Ehrenfeld, Edward R.

- Melnick, David W. Bates, Julia Adler-Milstein, and Rebecca G. Mishuris. Changes in clinician time expenditure and visit quantity with adoption of artificial intelligence-powered scribes: A multisite study. *JAMA*, 335(16):1408–1417, 2026. doi: 10.1001/jama.2026.2253. URL <https://doi.org/10.1001/jama.2026.2253>.
- E. Sanmark, V. Vartiainen, J. Sanmark, K. Wettin, L. Saari, and A. Entezarjou. Evaluation of an ai medical scribe after 236,153 notes generated across care levels in a european health system: Mixed methods retrospective observational study. *JMIR Medical Informatics*, 14:e90052, 07 2026. doi: 10.2196/90052.
- Louise Olsson, Marek Czajkowski, Rebecka Klang, Rolf Ahlén, and Lars Breimer. A systematic review comparing ambient scribes and conventional documentation methods in outpatient care. Technical Report 2026:86, Region Örebro län, Centre for Assessment of Medical Technology in Örebro (CAMTÖ), 2026. URL <https://www.regionorebrolan.se/siteassets/media/forskning/hta-camto/rapporter/rapporter-2026/2026.86-a-systematic-review-comparing-ambient-scribes-and-conventional-documentation-methods-in-outpatient-care.pdf>. Regional health technology assessment report; literature search through December 2025.
- Aisling Bracken, Sean Whelehan, Anita Rose Babu, Khalid Merghani, Eoin Sheehan, and Iain Feeley. Exploring the potential of ambient AI for inpatient documentation: A qualitative study with junior doctors. *Journal of Medical Systems*, 50:111, 2026. doi: 10.1007/s10916-026-02437-7. URL <https://doi.org/10.1007/s10916-026-02437-7>.
- V. Vallejo, J. Farreras, L. Valero, C. González-Grado, I. Peral, I. Cano, C. Launes, and Grupo de estudio LAIA-DOC. Escriba médico con ia en consultas pediátricas: Carga documental, calidad y límites de escalado [ambient ai scribe in pediatric outpatient visits: Documentation burden, quality, and scale-up limits]. *Medicina Clínica*, 166(8):107541, 2026. doi: 10.1016/j.medcli.2026.107541.
- Mathilde Vindstrup Østergaard. Noteless – ai-baseret ambient scribe i psykiatrisk klinik-praksis, 2026. URL <https://2026.e-sundhedsobservatoriet.dk/wp-content/uploads/sites/41/2026/05/B3-Mathilde-Vindstrup-Ostergaard.pdf>. Conference presentation abstract, E-sundhedsobservatoriet 2026.
- Kathrin Cresswell, Catharine Rose, Jessica Howdle, Lucas Martinus Seuren, and Robin Williams. The regional implementation of an electronic health record-integrated ambient scribe in primary and secondary care in england: Real-time qualitative evaluation. *JMIR Medical Informatics*, 14:e88472, 2026. doi: 10.2196/88472. URL <https://doi.org/10.2196/88472>.
- Lucas Martinus Seuren, Robin Williams, and Kathrin Cresswell. Beyond productivity: a NASSS-informed review of implementation risks and research priorities for ambient AI scribes in healthcare. *BMJ Digital Health & AI*, 2(1):e000091, 2026. doi: 10.1136/bmjdh-2026-000091. URL <https://doi.org/10.1136/bmjdh-2026-000091>.
- Aaron A. Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Balleca, Sarah B. Wilson Hannay, Kate Castilla, Cindy S. Lau, Patricia Kipnis, Vincent Liu, and Kristine Lee. Ambient artificial intelligence scribes: Learnings after 1 year and over 2.5 million uses. *NEJM Catalyst Innovations in Care Delivery*, 6(5), 2025. doi: 10.1056/CAT.25.0040. URL <https://doi.org/10.1056/CAT.25.0040>.
- Haute Autorité de santé and Commission nationale de l’informatique et des libertés. Accompagner le bon usage des systèmes d’intelligence artificielle en contexte de soins, February 2026. URL https://www.has-sante.fr/upload/docs/application/pdf/2026-03/accompagner_le_bon_usage_des_systemes_dintelligence_artificielle_en_contexte_de_soins_guide_has_cnildocument_de_travail.pdf. Document de travail, version du 16 février 2026.
- Läkemedelsverket. Läkemedelsverket granskar AI-assistenter i sjukvården, March 2026. URL <https://www.lakemedelsverket.se/sv/nyheter/lakemedelsverket-granskar-ai-assistenter-i-sjukvarden>. News release.
- R. Potts. Driving ai value: How healthcare cios can harness ambient scribes for roi. Gartner, 02 2026.

- 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. URL <https://doi.org/10.2196/86474>.
- Peter D. Stetson, Suzanne Bakken, Jesse O. Wrenn, and Eugenia L. Siegler. Assessing electronic note quality using the physician documentation quality instrument (pdqi-9). *Applied clinical informatics*, 3:164 – 174, 2012. URL <https://api.semanticscholar.org/CorpusID:1283186>.
- Erin Palm, Astrit Manikantan, Herprit Mahal, Srikanth Subramanya Belwadi, and Mark E. Pepin. Assessing the quality of ai-generated clinical notes: validated evaluation of a large language model ambient scribe. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025. ISSN 2624-8212. doi: 10.3389/frai.2025.1691499. URL <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1691499>.
- Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The jets benchmark of llm-as-judges as test-time scaling evaluators. *arXiv preprint arXiv:2504.15253*, 2025.
- OpenAI. Introducing GPT-5.4, March 2026. URL <https://openai.com/index/introducing-gpt-5-4/>.
- Anthropic. Introducing Claude Opus 4.6, February 2026. URL <https://www.anthropic.com/news/claude-opus-4-6/>.

A Related Work

The role and status of ambient scribes in healthcare systems A growing body of research has characterized the benefits of ambient documentation systems, common challenges in piloting and implementation, and patterns of adoption across healthcare markets. Recent randomized trials and multi-site evaluations suggest that ambient scribes can reduce documentation burden and improve aspects of clinicians’ well-being, although effects vary across products, users, and clinical settings [Olson et al. [2025], Pearlman et al. [2025], Lukac et al. [2025], Afshar et al. [2025], Chowdhury et al. [2026], Rotenstein et al. [2026]].

Evidence on the adoption and impact of ambient scribes in Europe remains more limited than in the United States. However, emerging studies from the Netherlands, Sweden, Ireland, Spain, and preliminary work from Denmark report similar benefits and challenges. These systems can reduce perceived documentation burden and improve clinician experience by freeing mental capacity and reducing perceived work pressure, while their effects on consultation time and operational efficiency remain variable [van Linschoten et al., 2026, Sanmark et al., 2026, Olsson et al., 2026, Bracken et al., 2026, Vallejo et al., 2026, Østergaard, 2026]. Implementation studies further show that ambient scribes alter clinical communication and redistribute documentation work. Sustained use thus depends on workflow fit, local adaptation, and organisational support [Cresswell et al. [2026], Seuren et al. [2026], Tierney et al. [2025]]. Reported challenges include the continued need for clinician review and correction, reduced performance in complex or multilingual consultations, technical and EHR integration issues, and uneven sustained adoption following initial pilots. At the organisational level, successful scaling also requires effective onboarding, workflow redesign, demonstrable financial returns, and increasingly mature approaches to data governance, clinical safety, and regulatory compliance [Haute Autorité de santé and Commission nationale de l’informatique et des libertés, 2026, Läkemedelsverket, 2026, Potts, 2026].

Comparative studies of ambient scribes Prior US-centric studies have compared commercial ambient scribes in provider-specific implementation settings focused primarily on time-saving and stress-reduction outcomes, pointing to quality issues but without going into detailed comparisons [Lukac et al., 2025, Chowdhury et al., 2026]. Comparisons using shared encounters and direct note-quality measures report substantial variability. Anderson et al. [2025] compared five ambient-scribe platforms on shared simulated encounters; Fox et al. [2026a] audited three commercial scribes, though differences in input modality and template contributions limit what can be inferred about the generation pipelines; and Draper et al. [2025] compared seven scribes on eight consultations, finding summarisation accuracy was generally high, omissions were the dominant error type, and error frequency varied widely across scribes.

Evaluation of Clinical Documentation Systems Evaluating generated clinical documentation requires assessing whether a note preserves important information, remains supported by the encounter, and presents that information usefully. Existing frameworks such as SCRIBE combine automated metrics, simulation, clinician review, and LLM-based evaluation [Wang et al., 2025, Taylor et al., 2026]. Omissions are a central challenge in evaluating generated clinical notes: information missing from a note can escape evaluators that primarily inspect the statements it contains, thus omission-focused work provides a complementary perspective: MED-OMIT decomposes dialogue into facts and assesses omitted information through its relevance to a simulated differential diagnosis [Schumacher et al., 2025]. More recent experiments show that explicitly enumerating source facts before checking the note can improve omission detection over whole-note judgments, although detection remains imperfect and calibration does not necessarily transfer to vendor-generated notes [Fox et al., 2026b].

Structured quality rubrics provide another complementary approach. Based on the initial Physician Documentation Quality Instrument (PDQI) introduced by Stetson et al. [2012] [Palm et al., 2025], the Provider Documentation Summarization Quality Instrument (PDSQI-9) was developed and validated for LLM-generated summaries of clinical records [Croxford et al., 2025a]. Subsequent work evaluated LLM judges against physician assessments using this instrument, supporting the potential for scalable, multidimensional assessment without establishing exhaustive clinical-error detection [Croxford et al., 2025b].

B Symphony for Text Generation

Text Generation API endpoints Symphony exposes clinical text and document generation via several API endpoints spanning general LLM inference, clinical fact extraction (FactsR [Hansen et al., 2025]) and guided document synthesis, and working in tandem with Symphony for speech-to-text [Nix et al., 2026], Symphony for medical coding [Edin et al., 2026], and agentic orchestration of multi-step clinical workflows. The guided document synthesis API accepts multiple forms of clinical context and return schema-controlled, structured clinical documents based on fully API-configurable templates. This modular architecture enables straightforward customization and orchestration of document-generation workflows, while supporting integration across any form factor. Rather than functioning as a standalone point solution, ambient documentation can be embedded directly into larger automated workflows supporting downstream actions.

Clinical context and fact extraction Text Generation supports several input representations, including conversation transcripts, extracted clinical facts, clinician notes taken during the encounter in the form of dictations or free-form clinical text, existing medical documents, and combinations of these sources. Clinical source material can be passed directly to document synthesis or first transformed into structured clinical facts using FactsR.⁹ The latter provides an intermediate representation between the source transcript from the consultation and the final note, allowing clinically relevant propositions to be identified and reviewed.

Guided document synthesis Documents are generated through guided, template-controlled synthesis. The generation layer receives the selected source context together with a template and produces a structured clinical document rather than unconstrained free-form text. Output can range from conventional narrative sections to lists, typed fields, arrays, nested objects, and other representations suitable for direct integration into downstream systems or EHR fields.

Templates and sections Templates are first-class, UUID-addressable resources composed of reusable, typed sections with defined output schemas. Rather than representing a document template as one monolithic prompt that is difficult to debug and calibrate, each section exposes several narrower control surfaces that separately specify what information should be included or excluded, the desired voice and writing style, additional synthesis instructions, and the required output format. This decomposition separates *what* should be documented, *how* it should be expressed, and *how* it should be structured, allowing each dimension to be isolated, tested, and tuned independently.

Sections can be reused and inherited across templates, allowing organizations to derive specialty-, department-, or use-case-specific variants from common base sections while changing only the instructions that differ. Improvements to a shared base section can therefore propagate across dependent templates without requiring equivalent changes to hundreds of independently maintained prompts. Corti additionally provides a curated library of clinically pre-configured “Corti Standards” templates and sections that can be combined directly, partially customized, or replaced with organization-specific definitions. Together with template versioning and publishing, this provides a governed and programmatically reproducible mechanism for controlling clinical documentation behavior. All template configuration functionality is available directly through an API and as a user-interface through the Corti Console.¹⁰

Note-generation pipeline used in these experiments In the experiment below, notes are generated through the full pipeline: FactsR first extracts facts from the transcript, after which both are passed to guided document synthesis together with any additional context. This reflects a typical real-time workflow in which transcripts and facts are generated during the encounter and subsequently used for note generation. Although FactsR supports clinician review, we use the extracted facts without modification to ensure a fair comparative analysis.

C System configurations

⁹<https://docs.corti.ai/textgen/factsr>

¹⁰<https://console.corti.app/>

Table 4: Configuration of the three clinical note-generation systems evaluated in the benchmark. All systems were provided with pre-transcribed consultation content; speech-to-text performance was therefore outside the scope of the comparison.

Parameter	Corti	Heidi	Tandem Health
Access method	Text Generation API	End-user web interface	End-user web interface
Input	Pre-transcribed consultation transcript; additional extracted clinical facts available to the generation pipeline	Pre-transcribed consultation transcript provided as context	Pre-transcribed consultation transcript provided as context
Speech-to-text evaluated	No	No	No
Markets / languages	UK English, German, Danish	UK English, German, Danish	UK English, German, Danish
Clinical note templates	SOAP and GP-SOAP configurations	SOAP configurations	SOAP configurations
Template source	Corti Standards and evaluation-specific configurations	Heidi defaults for each market, adapted where necessary	Tandem Health defaults where available; templates created from scratch where no suitable default was available
Template representation	Typed, section-based templates with explicit section-level instructions for content, style, synthesis, and output structure	Editable template structure with vendor-provided defaults and visible prompt configuration	Section- and heading-based templates; some system-provided section instructions are not exposed to the evaluator
Template harmonization	Section structure normalized to the common benchmark taxonomy	Vendor templates minimally adapted to the common benchmark taxonomy	Vendor templates minimally adapted to the common benchmark taxonomy
Prompt transparency	Section-level generation instructions available to the evaluator	Template prompt visible through the product’s Structure view	Underlying system-provided section prompts not fully visible
Generation configuration	Evaluation configuration held fixed across all cases within each condition	Scribe: Simple; Voice: Goldilocks (default)	Assessment: “Brief assessment & Plan” (default); Level of detail: Standard (default)
Configuration policy	Fixed for each experimental condition	Vendor defaults retained except for required template normalization	Vendor defaults retained except for required template normalization
Evaluation date	September 2026	September 2026	September 2026

C.1 Template Configuration Details

Table 5: Heidi SOAP template configuration by country.

Country	Default available	Adjustments
UK	Yes	Moved prompts of separate Past Medical History section into Subjective section to normalise to 4 SOAP sections
Germany	Yes	Adjusted German SOAP-Notiz by moving all of Past Medical History (Vergangene Krankengeschichte) under Subjective section to normalise to 4 SOAP sections. Removed (S), (O), (A), (P) from headings to normalise.
Denmark	Yes (English prompts)	Exact same adjustments as with UK SOAP

Table 6: Tandem Health SOAP template configuration by country.

Country	Default available	Adjustments
UK	No	Created from scratch, headings only
Germany	Yes	Based on standard German EHR template (SOP), separated “Beurteilung” (Assessment) from “Plan” to normalise to 4-section SOAP
Denmark	Yes	Based on SOAP for Almen Praksis, separated “Vurdering” (Assessment) from Plan to normalise to 4-section SOAP

Table 7: Corti SOAP template configurations

Note: You can access Corti templates and their configurations via Corti Console or the public API to retrieve the full prompt configuration for each uuid listed below.

Country	Default available	Adjustments
UK	Yes SOAP: f901c06a-70db-59f6-8d0a-0a4bef4b8c77 succinct SOAP : b9f26972-f8ff-5a6d-aa75-e9cd43d783fa	none
Germany	Yes SOAP: 549b5f41-aed2-51fe-b0a2-f963bb7d6555 succinct SOAP: d85de77b-e5e9-5b5c-976a-ea21fc0e0b1c	none
Denmark	Yes SOAP: dc601541-c241-5ee5-90ad-b39bf53516e4 succinct SOAP: c85d2aaa-26ee-5a81-986f-a4cbdeaed9eb	none

D Template Architecture Differences

Table 8: Template architecture and configuration transparency across the evaluated systems.

Property	Corti	Heidi	Tandem Health
Template representation	Typed, section-based templates composed of reusable document sections	Editable template structure with section headings and associated prompting	Primarily section- and heading-based templates
Content control	Explicit section-level instructions define content to include or exclude	Content instructions can be specified within the editable template prompt	Template headings define intended content; additional custom instructions can be added
Style and synthesis control	Separate controls for writing style, synthesis instructions, and output representation	Style and content instructions are expressed within the template prompt	Custom prompting and user- or clinic-level preferences can influence generated output
Output structure	Section-specific, schema-controlled outputs, including typed structured fields	Template structure determines the organization of the generated document	Template headings determine document structure; selected fields can use structured output types
Prompt transparency	Section-level content, style, synthesis, and formatting instructions are exposed	Template prompt can be inspected through the Structure view	System-provided section-level instructions are not fully exposed
Reuse and inheritance	Sections are reusable and versioned; templates can inherit from shared base sections	Vendor defaults can be copied, modified, or replaced with custom templates	Vendor defaults can be modified or new templates can be created
Vendor-provided templates	Curated Corti Standards library of clinical templates and reusable sections	Vendor-provided defaults plus a community library of clinician-created templates	Market- and specialty-specific vendor defaults where available
Localization	Templates can be configured as locale-specific variants	Market-specific defaults are available, although prompt localization varies by market	Market-specific defaults are available, with template availability varying by market

E Evaluation metrics

E.1 Entailment-based evaluation metrics

In this work, we frame the semantic alignment metrics of Hansen et al. [2025] as a textual entailment task and use an LLM as the entailment judge. Given a segmented *source* (the premise) and a set of segmented *hypotheses*, the judge decides, for each hypothesis independently, whether it is entailed by the source. Because the judge reasons only over the sentences it is given and may not introduce outside facts, the result is a *directional* semantic relation between the transcript T , the reference note D_R , and the generated note D_G , rather than a learned proxy of preference.

Formally, the judge takes a premise segmented into sentences $P = (p_1, \dots, p_m)$ and hypotheses $H = (h_1, \dots, h_n)$, and assigns every hypothesis h_j a label

$$\ell(h_j) \in \{\text{entailed}, \text{partially entailed}, \text{not entailed}, \text{not applicable}\},$$

citing the source sentences it relied on. A hypothesis is `entailed` when every claim it makes is stated in, or follows by deduction from, one or more cited source sentences; `partially entailed` when its main claim is supported but it adds a detail the source does not support (e.g., an unstated dose, frequency, duration, or laterality) or only some of its claims hold; `not entailed` when it is unsupported or contradicted; and `not applicable` for section headers and non-clinical boilerplate. A hypothesis may be supported jointly by several source sentences, and a source sentence may support several hypotheses.

Each label is mapped to a weight, `entailed` = 1, `partially entailed` = $\frac{1}{2}$, `not entailed` = 0, and a metric is the mean weight over the applicable hypotheses:

$$\text{score}(H) = \frac{\sum_{j: \ell(h_j) \neq \text{n/a}} w(\ell(h_j))}{|\{j : \ell(h_j) \neq \text{n/a}\}|},$$

where `not applicable` hypotheses are excluded from both numerator and denominator, and the score is defined as 1 when every hypothesis is `not applicable` so that boilerplate-only text is not penalized. The intermediate `partially entailed` label makes this finer-grained than a binary supported/unsupported decision.

The three metrics differ only in which document acts as the premise and which supplies the hypotheses:

1. **Completeness:** premise D_G , hypotheses the segments of D_R , giving the proportion of D_R entailed by the generated note. A score of 100% means all meaning in D_R is captured; 0% means none is.
2. **Conciseness:** premise D_R , hypotheses the segments of D_G , giving the proportion of the generated note entailed by D_R , reflecting clinical relevance. A score of 100% means all generated content is relevant; 0% means none is.
3. **Groundedness:** premise T , hypotheses the segments of D_G , giving the proportion of the note entailed by the transcript. This differs from a hallucination rate because unsupported sentences may also stem from entailment-judge or segmentation error rather than a genuine fabrication.

Completeness and conciseness thereby play the roles of recall and precision of the generated note against D_R , while groundedness measures support in T .

E.2 Pairwise Evaluation

To capture fine-grained differences in quality between systems, we evaluate relative note quality using model-based pairwise measures. An LLM compares two notes generated from the same source transcript and determines which note better satisfies a set of predefined quality dimensions. This provides a measure grounded in differences between the notes, as opposed to global calibrated score, avoiding challenges associated with numerical scores and LLM-based judges [Zhou et al., 2025]. Additionally, being a reference-free measure, it provides a scalable means of evaluation on a much larger and more diverse collection of clinical input, than that by the scarce availability of annotated clinical notes.

For rubric dimensions, we adopt the validated quality measures for evaluating the quality of clinical records, the *Provider Documentation Summarization Quality Instrument* (PDSQI-9) [Croxford et al., 2025a,b]. See Table 9 for a definition of each dimension. For each pair, the judge receives the source transcript, two notes, all dimension definitions, and returns one of three outcomes for each dimension—Note A, Note B, or Tie—together with a brief justification. The following dimensions are included: *Accurate*, *Thorough*, *Useful*, *Organized*, *Comprehensible*, *Succinct*, *Synthesized*, and *Stigmatizing*. We exclude the *Cited* attribute because the evaluated systems do not produce provenance citations.

We add an aggregate *Overall* dimension based on the majority vote of the eight retained dimensions. For each of the eight dimensions, the judge prefers system A, prefers system B, or records a tie. A system receives the overall preference if it is preferred on at least five dimensions; otherwise, the overall result is a tie. We recognise that note quality is highly subjective, and that there is no evidence that dimensions are linearly correlated with preference, however, this provides a simple but expressive summary statistic to guide intuition.

Table 9: Dimensions of the Provider Documentation Summarization Quality Instrument (PDSQI-9) Croxford et al. [2025a].

Dimension	Description
Cited	Claims are accompanied by appropriate citations to the source documentation.
Accurate	Information is factually correct and free from fabrication, falsification, or other incorrect content.
Thorough	Important clinical information is sufficiently complete, without clinically relevant omissions.
Useful	Included information is relevant and useful to the clinician receiving the summary.
Organized	Information is structured coherently so that the clinical course and relevant information can be followed easily.
Comprehensible	The summary is clear and unambiguous and uses understandable language and terminology.
Succinct	Information is communicated concisely, avoiding unnecessary repetition or redundant content.
Synthesized	Information is integrated into a coherent clinical synopsis that reflects appropriate inference and medical reasoning.
Stigmatizing	The summary is free from stigmatizing language about the patient.

To mitigate positional bias, a pair is evaluated twice. In the first pass, system outputs X and Y are presented as Note A and Note B, respectively; in the second pass, their positions are reversed. We then map both responses from positional labels back to the canonical system identities, yielding $z_1, z_2 \in \{X, Y, T\}$, where T denotes a tie. The final outcome is defined conservatively as

$$c(z_1, z_2) = \begin{cases} X, & z_1 = z_2 = X, \\ Y, & z_1 = z_2 = Y, \\ T, & \text{otherwise.} \end{cases} \quad (1)$$

A note is declared *preferred* only if both judgments select it. Any disagreement resolves as a tie. While this prevents position-sensitive judgments from being counted as a win, it does not remove other sources of model-judge bias.

For a system under review X , we summarize the adjudicated outcomes using a tie-adjusted preference score

$$\text{PS}(X) = 100 \frac{N_X + \frac{1}{2}N_T}{N_X + N_Y + N_T}, \quad (2)$$

where N_X , N_Y , and N_T are the numbers of adjudicated wins for X , wins for Y , and ties, respectively. A score of 50 is neutral; values above 50 favor X , and values below 50 favor Y .

To estimate 95% confidence intervals for the preference scores, we use a non-parametric bootstrap over paired notes. Within each bootstrap replicate, ordered pairs are sampled with replacement, and the preference score is recomputed from the resulting win, loss, and tie outcomes. All dimension-level outcomes belonging to an encounter are resampled together, preserving their within-encounter dependence. For results pooled across datasets, resampling is performed separately within each dataset to preserve the original dataset sizes. The reported interval is defined by the 2.5th and 97.5th percentiles of the scores obtained from 10,000 bootstrap replicates.

E.3 Judge Sensitivity

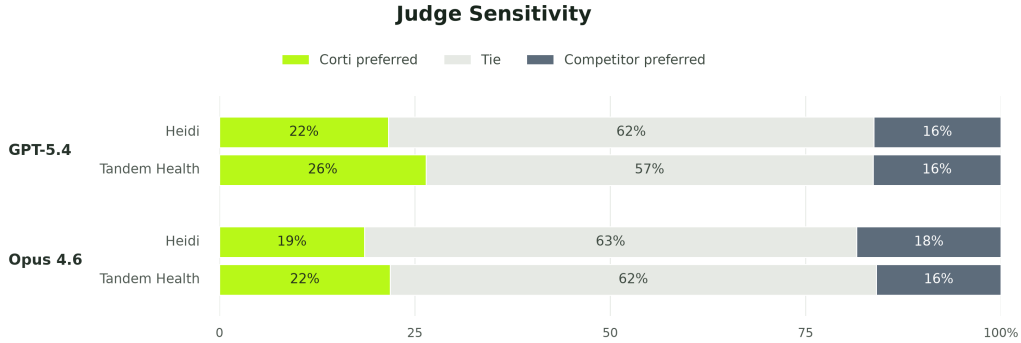


Figure 4: Judge sensitivity under GPT-5.4 and Opus 4.6 across all datasets and dimensions.

Adopting LLM-as-a-judge measures, inherently adopts the bias and preferences of the chosen underlying judge model. We conduct a simple sensitivity analysis of preference scores to identify potential variation across LLM families. We consider OpenAI’s GPT-5.4 [OpenAI, 2026] and Anthropic’s Opus 4.6 [Anthropic, 2026]. We recompute preference scores anew, as described in §E.2, and display "Overall" preference scores in Figure 4. Both models exhibit similar preferences distributions. Opus 4.6 more frequently scores pairs as ties than GPT-5.4, while both models generally lean towards Corti as being the more preferred output *Overall*. The results show little difference between model families, indicating low judge sensitivity for the proposed pairwise comparison method.

F System Latency Extended

Figure 5 presents note-generation times for each system across the four datasets. Corti displays the fastest runtime across systems and datasets, with lower median, fewer outliers, and narrower interquartile range. These distributions complement the mean runtimes in Table 3. We emphasize that Heidi and Tandem Health were measured through their browser interfaces, therefore, distributions should be interpreted as indicative

G Targeted clinician validation of the pairwise judge

Clinicians agreed with the LLM judging, when shown its verdict and reasoning, in most assessments (207 of 240 samples, i.e. 86.3%), with agreement strongest on accuracy, thoroughness, and usefulness and weakest on synthesis (Table 10). At the case-dimension level, a majority of clinicians supported the LLM judgment in 75 of 80 comparisons (93.8%), including unanimous agreement in 52 comparisons (65.0%).

Overall Gwet’s AC1 was 0.694 indicating substantial overall agreement. Dimension-level AC1 was highest for accuracy (0.929), thoroughness (0.837), and usefulness (0.827), and lowest for synthesis (0.080). Nominal Krippendorff’s alpha was 0.020 overall, consistent with the strong imbalance in the binary agree/disagree response distribution.

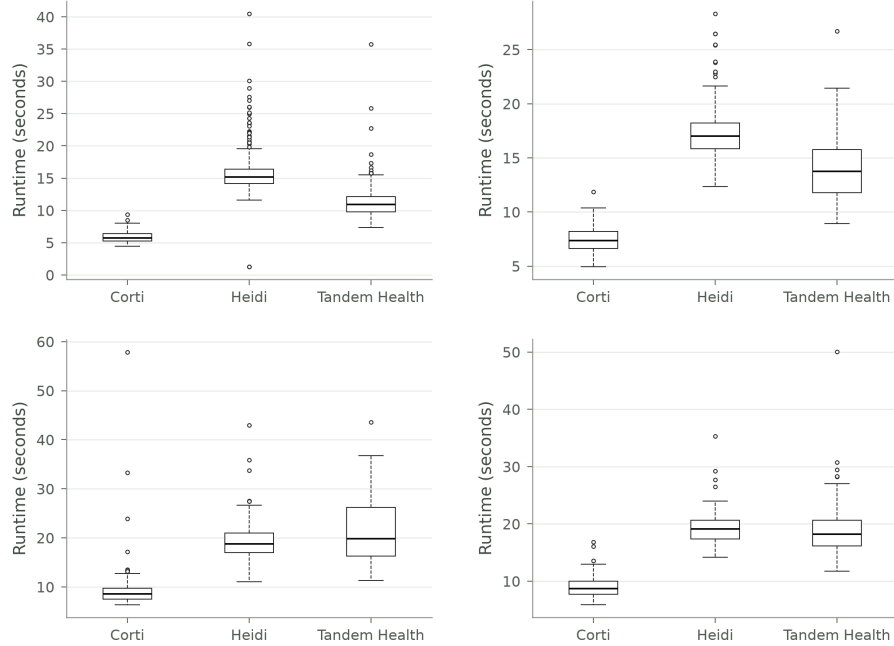


Figure 5: Comparison of the latency speed plots for ACI-BENCH (top left), MedConv EN (top right), DA (bottom left) and DE (bottom right).

Table 10: Inter-clinician agreement with the LLM judge across PDSQI dimensions.

Dimension	Agreement	Pairwise agreement	Krippendorff’s α	Gwet’s AC1
Overall	207/240 (86.3%)	0.767	0.020	0.694
Accurate	29/30 (96.7%)	0.933	0.000	0.929
Thorough	27/30 (90.0%)	0.867	0.284	0.837
Useful	26/30 (86.7%)	0.867	0.442	0.827
Organized	24/30 (80.0%)	0.600	-0.208	0.412
Comprehensible	26/30 (86.7%)	0.733	-0.115	0.653
Succinct	24/30 (80.0%)	0.667	-0.007	0.510
Synthesized	21/30 (70.0%)	0.467	-0.228	0.080
Stigmatizing	30/30 (100.0%)	1.000	—	1.000

Note. Agreement denotes the proportion of clinician evaluations that agreed with the LLM judgment. Krippendorff’s α and Gwet’s AC1 were calculated on the binary agree/disagree ratings. Krippendorff’s α was undefined for the stigmatizing-language dimension because all ratings were identical.

Qualitative post-annotation interviews highlighted that the annotation task and UI was very clear overall, with dimensions varying in difficulty to assess. All clinicians stated they had a very clear favourite note for each case, however the driver of this clear preference differed between a clear preference for succinctness for one, a more thorough fluent sentence writing style preference for another, and a bias towards succinctness but with critical flaws in the note quickly determining a winner for the third clinician.

H Limitations

The following limitations should be noted.

1. **Template prompt transparency.** Corti exposes all section-level prompts fully. Heidi exposes the underlying prompt in a Structure-view. Tandem Health does not expose its section-level prompts; templates created from headings only are assumed to be outfitted with hidden Tandem Health-authored prompts. This asymmetry means that for Tandem Health,

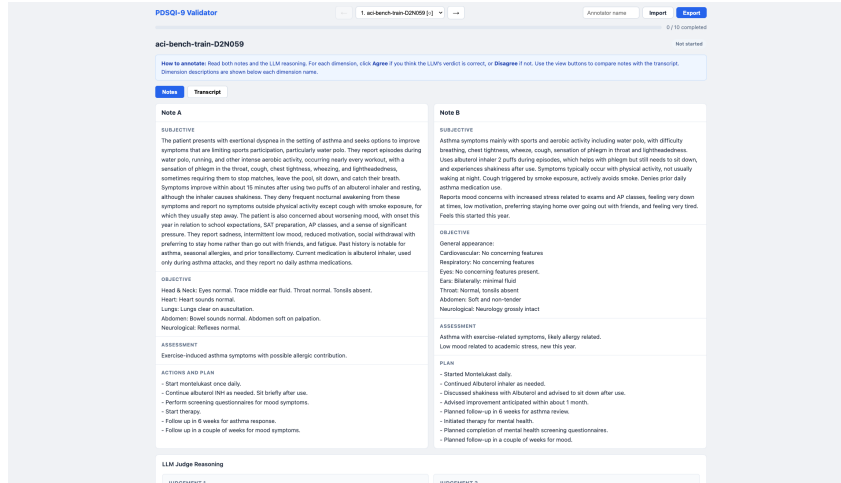


Figure 6: Validation tool showing compared system notes side-by-side, flipped randomly

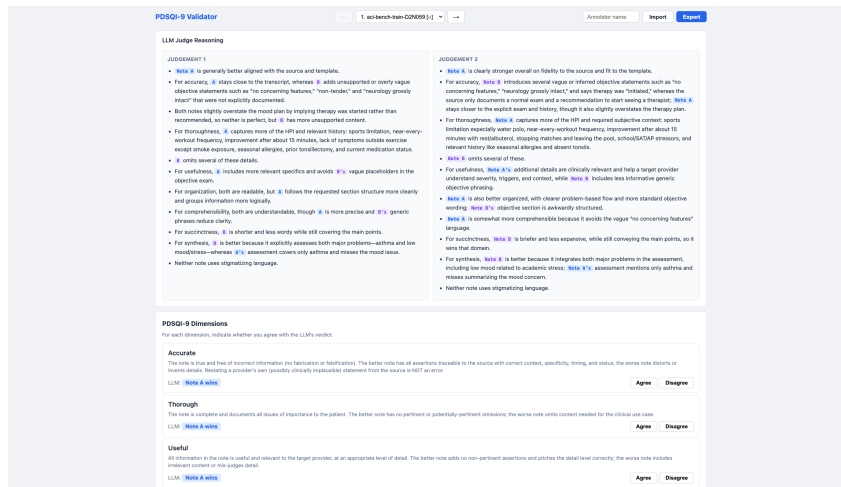


Figure 7: Validation tool showing LLM judgements and dimensions to validate

the effective prompt is not fully known and cannot be directly compared to the prompts used for Corti and Heidi. Differences in output quality may therefore reflect differences in hidden prompt engineering rather than in pipeline or model capability, and this could favour Tandem Health as easily as it disadvantages it. Tandem Health also had no suitable vendor-provided UK template, so the English comparisons evaluate it under a template we built from section headings alone.

- Model and infrastructure differences.** The three systems use different underlying LLMs and cloud infrastructure. Corti uses its own models accessible via API. Heidi’s subprocessors indicate GCP hosts “many of our AI models” alongside AWS and Azure. Tandem Health’s subprocessor list names Microsoft Azure for “AI models and related processing services” and ElevenLabs for speech-to-text. The systems’ LLMs, model versions, and inference configurations are not documented and may change over time. Model, prompt, template, and serving infrastructure vary together across the three systems, so observed differences are attributable to the systems as configured and served, not to any single component.
- Clinician validation measured ratification, not independent preference.** Clinicians were shown the LLM judge’s declared winner together with its reasoning trace and asked only to agree or disagree, rather than recording an independent preference for later comparison. This anchors the annotator on the judgment under test, so the reported agreement rate and Gwet’s AC1 should be read as upper bounds on judge–clinician concordance.

4. **System-level configuration parameters.** Heidi and Tandem Health both expose user-facing settings that influence output style and length (Heidi: voice modes brief/detailed/goldilocks, Simple vs. Pro scribe; Tandem Health: “Brief assessment & Plan” toggle, “Level of detail” beta with Standard/Concise/Comprehensive). These were held constant at their respective defaults, but they represent system-specific parameters that are injected together with a template. In Corti, equivalent voice, style and length controls are explicitly part of the section prompting instructions itself for full transparency and control.
5. **Language and market localisation.** Heidi’s Danish templates shipped with English-language prompts despite Danish headings. Tandem Health’s Danish SOAP was based on a Danish-specific template (Almen Praksis). Corti’s templates were configured with per-locale variants. The degree of native-language prompt localisation varies and may affect output quality in non-English markets, particularly Danish.

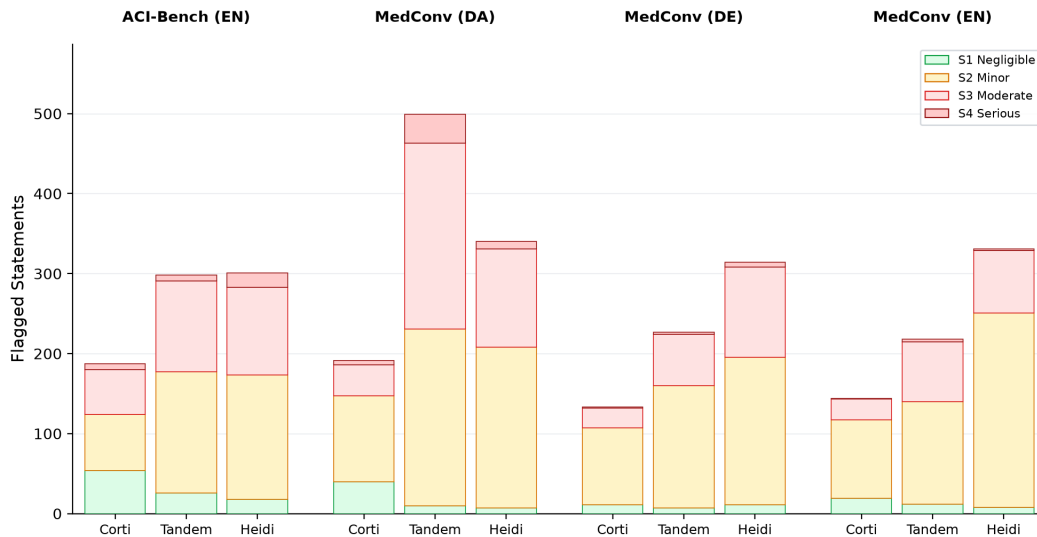
I Characterization of ungrounded statements

As a preliminary analysis of groundedness failures, we used an LLM (Corti-S1¹¹) to categorize statements flagged as partially or not entailed and assign a severity level. Across all 412 encounters, Corti accumulated 655 flagged statements (160 at severity S3+), compared with 1,242 (534 S3+) for Tandem Health and 1,286 (459 S3+) for Heidi. Fully unsupported statements were least frequent for Corti (23), followed by Heidi (83) and Tandem Health (149); most flagged content across all systems was only partially entailed, where the core claim was supported but an unstated detail had been added.

The dominant failure mode across all three systems was *specificity overreach*, such as adding a medication route or laterality not present in the source. This accounted for 22%, 26%, and 40% of flagged entries for Corti, Tandem Health, and Heidi, respectively. Many of these cases were low severity because the added detail was plausibly correct—for example, adding PO for a medication available only orally—but still failed strict entailment because the detail was not explicitly supported by the transcript. System-specific patterns also emerged: Tandem Health had the highest rate of *fabrication from silence* (195 entries, versus 25 for Corti and 118 for Heidi), while Heidi showed more *terminology substitution* (156 entries, versus 55 for Corti).

Because both failure-mode classification and severity scoring were LLM-assigned and the rubric has not been independently clinically validated, this analysis should be interpreted as an indicative characterization of the errors captured by the groundedness metric rather than a confirmed clinical error audit.

Figure 8: Flagged Statements, partially or fully ungrounded, by Dataset x Severity



¹¹<https://docs.corti.ai/models>

Table 11: Examples of ungrounded statements by system and severity (S1 negligible, S2 minor, S3 moderate, S4 severe). All examples are drawn from ACI-Bench; encounter IDs omit the aci-bench- prefix. Justifications are abridged from the LLM classifier output.

Sev.	Flagged statement	Category and justification
Corti		
S1 test1-D2N108	Rx clindamycin 400 mg PO twice daily for 7 days.	<i>Added route / frequency.</i> The clinician said “you take that,” implying oral administration; adding “PO” would not mislead a reader.
S2 test1-D2N107	Rx physical therapy with home stretches and exercises.	<i>Specificity overreach.</i> The source supports PT with stretches and exercises but does not specify “home.” Clinically implied and unlikely to change management.
S3 test1-D2N109	Air splint applied. Crutches provided. Non-weightbearing until radiograph result.	<i>Contingent to definitive.</i> The clinician describes a plan (“I’m gonna put you in an air splint”), but the note records it as completed, misrepresenting what was done during the visit.
S4 valid-D2N077	Ulnar styloid fracture present.	<i>Fabrication from silence.</i> The source mentions ulnar styloid fracture only in a general description of Colles’ fractures, not as a finding for this patient. The note asserts a fracture that was not identified.
Tandem Health		
S1 test1-D2N106	Start Montelukast 10 mg, 1 tablet once per day.	<i>Specificity overreach.</i> “1 tablet” is implied by a 10 mg prescription but is more specific than the source; it would not change management.
S2 valid-D2N081	Denies mucus, shortness of breath, nausea, vomiting, sensation of food getting stuck.	<i>Specificity overreach.</i> The patient’s answer “more dry” supports a dry cough but is not an explicit denial of mucus.
S3 test1-D2N123	Thyroid: Thyroid gland symmetric, non-enlarged, smooth, without nodules or tenderness.	<i>Specificity overreach.</i> The source states only that the thyroid “feels normal”; the listed findings are not stated and overstate the thoroughness of the exam.
S4 train-D2N038	No hepatosplenomegaly.	<i>Temporal distortion.</i> The clinician corrects the initial “no hepatosplenomegaly” to “splenomegaly”; the note retains the superseded negative finding.
S4 train-D2N047	No chest pain or shortness of breath.	<i>Fabrication from silence.</i> The clinician asks about these symptoms but the patient never answers; the negatives are unsupported in a patient with known coronary artery disease.
Heidi		
S1 test1-D2N111	Continue Lisinopril 20 mg daily.	<i>Added route / frequency.</i> Lisinopril is dosed once daily by convention; adding “daily” would not mislead a clinician.
S2 test1-D2N109	Experienced immediate soreness, swelling, and pain, making it difficult to walk.	<i>Specificity overreach.</i> “Immediate” onset is not stated in the source. Plausible, but adds unsupported temporal detail.
S3 test1-D2N110	Right foot examination: 1×2 cm circular wound on the dorsal aspect of the lateral right foot, proximal to the fifth MT.	<i>Wording mismatch.</i> The source gives 1×2 inches (about 2.5×5 cm); the note’s unit change misstates the wound size.
S3 test1-D2N112	Lower back pain for one month, chronic history of back pain for 30 years.	<i>Specificity overreach.</i> The source mentions a fall 30 years ago and that the back “has been bothering me for a long time,” not a continuous 30-year history.
S4 test1-D2N119	Investigations planned: None.	<i>Fabrication from silence.</i> The source describes reviewing pulmonary function tests and planning an asthma action plan; the note asserts an absence of workup.
S4 test3-D2N187	No PE changes in the fovea.	<i>Other.</i> The source states “or pe changes in the fovea” for the left eye, indicating RPE changes are present; the note asserts the opposite.