Klara: Medical Speech Recognition for Real-World Clinical Environments

Frédéric Odermatt **Florence Lentner** **Andrin Benz**
Colin Kolbe **Nicolas Gürtler** **Jannis Schönleber**
44ai Labs
{frederic.odermatt, florence.lentner, andrin.benz,
colin.kolbe, nicolas.guertler, jannis.schoenleber}@44ai.ch

Abstract

General-purpose speech recognition systems fall short in German-speaking clinical practice, where dense clinical terminology and high safety standards meet a multitude of dialects and a growing share of non-native speaking medical staff. We present Klara, a medical transcription model for German and its dialects. Klara builds on a strong multilingual speech recognition backbone and is trained on clinical dictations and ambient consultations covering Swiss, Austrian, and German dialect groups, non-native speech, and more than 150'000 medical terms. On `44ai-clinical-de`, a 12-hour multi-dialect clinical dictation set with held-out clinical topics, Klara has the highest word accuracy of twelve evaluated systems (92.7% vs. 88.6% for the strongest competitor) and the highest medical term accuracy (90.6% vs. 87.9%). The margin is largest on dialectal speech: on Swiss German speakers, Klara reaches 90.7% word accuracy against 81.7%. On the German portions of Corti's MedTerm and MedDictate benchmarks, Klara is within two points of the best word accuracy. We release the validation set and our scoring code, including all normalization rules.

1 Introduction

Clinical documentation consumes a large share of clinicians' working time, and speech is becoming its primary interface: clinicians dictate findings, letters, and notes (Goss et al., 2019; Hodgson et al., 2017), and increasingly rely on ambient transcription of full consultations, which reduces consultation time, cognitive load, and documentation effort (van Buchem et al., 2024). Transcription accuracy remains the limiting factor for adoption (Ha et al., 2025; Quiroz et al., 2019), and errors are not cosmetic: a swapped medication name, a wrong laterality, or a missed negation changes the clinical meaning of a document (Zhou et al., 2018).

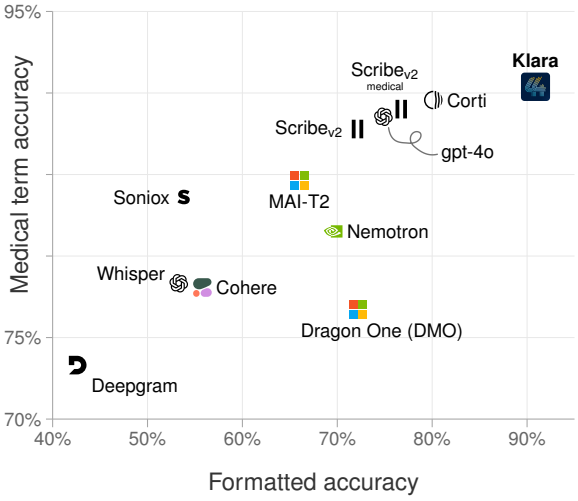


Figure 1: Formatted accuracy (100 − fWER) versus medical term accuracy, combined over `44ai-clinical-de`, MedTerm, and MedDictate and weighted by audio duration (Tables 1 and 3).

A clinically usable model must solve three problems at once. It must recognize dense, safety-critical terminology that is rare in general speech corpora. It must handle the speech clinicians actually produce, which in much of the German-speaking world is not standard German: Swiss German diglossia requires transcribing dialect into standard written German (Plüss et al., 2023; Dolev et al., 2024), Austrian and regional variation is pervasive, and a growing share of medical staff speak German as a second language. And it must be robust to everyday clinic acoustics.

Existing work addresses these axes in isolation. General-purpose models such as Whisper (Radford et al., 2023) degrade substantially on Swiss German and regional dialects (Dolev et al., 2024; Blaschke et al., 2025). Dialect adaptation has improved general-domain speech (Sicard et al., 2023;

`44ai-labs/44ai-klara-evaluation`
`44ai/44ai-clinical-de-validation`

Timmel et al., 2025) but not clinical terminology. Strong medical systems exist for English (Nix et al., 2026; Adedeji et al., 2024), while for German a benchmark of 29 systems finds that none reliably handles medical terminology in dialectal speech (Schuster et al., 2026). General-purpose systems such as gpt-4o-transcribe (OpenAI, 2025), ElevenLabs Scribe v2 (ElevenLabs, 2026a), and Cohere Transcribe (Cohere Labs, 2026) are accurate on broad-domain speech but not built for German clinical practice.

We present Klara[1], a medical transcription model for German and its dialects, trained on clinical dictations and ambient consultations that cover more than 150'000 medical terms, Swiss, Austrian, and German dialect groups, and non-native speech, combining real recordings with synthetic data that injects rare terminology into realistic clinical narratives. Our contributions are:

1. **A dialect- and accent-robust German medical transcription model** with the highest formatted and medical term accuracy of twelve evaluated systems over three clinical test sets (Figure 1), and the highest word accuracy on a multi-dialect clinical validation set (Table 2).

2. **A clinical validation set:** We release `44ai-clinical-de`, 12 hours and 4'331 dictations across nine German dialects and accents, with clinical topics held out from training (Section 3).

3. **Open scoring code:** We release our evaluation harness with all normalization rules, so every number in this paper can be reproduced (Section 4.2).

2 Training Set

Klara is trained on a large corpus of German clinical speech paired with transcripts, built around the three axes on which general-purpose systems fail: dialectal and accented speech, dense medical terminology, and realistic documentation modes.

2.1 Setup

Each example pairs an audio segment with a standard-German reference transcript and metadata: dialect, documentation mode, speaker gender, topic, and the medical terms in the reference. Two documentation modes are covered:

[1]Klara is served in production as part of the Haidy® medical intelligence platform.

INTENDED RECORD (Zürichdeutsch dictation)
. . . zeiged düütlech, dass in dere spezielle Patientenkohorte en funktionierendi Integrierti Versorgig essenziell isch. Mir rege daher die ziitnah. . .

ELEVENLABS SCRIBE V2 (API DEFAULT)
. . . zeigen tydlich, dass in dere spezielle Patientenkohorte äh funktionierende integrierte Versorgung essentiell ist. Mir rege daher die zeitneu. . .

KLARA (OURS)
. . . zeigen **deutlich**, dass in **dieser speziellen** Patientenkohorte **eine** funktionierende integrierte Versorgung essenziell ist. **Wir regen** daher die **zeitnahe**. . .

Figure 2: **The task** A Zürichdeutsch clinical dictation from `44ai-clinical-de`: the clinician speaks dialect, the record requires standard German. At its API default, ElevenLabs Scribe v2 leaves Swiss German grammar in the transcript (*Mir rege* for *Wir regen*, *dere spezielle* for *dieser speziellen*) and garbles words (red). Klara renders the intended standard German (teal). Outputs are verbatim. Casing, punctuation, and folded spelling variants are not marked.

Dictation Single-speaker dictations mirroring radiology and letter dictation: dense terminology, spoken formatting commands, little conversational structure.

Ambient Multi-speaker consultations with turn-taking, disfluencies, and colloquial framing around clinical content.

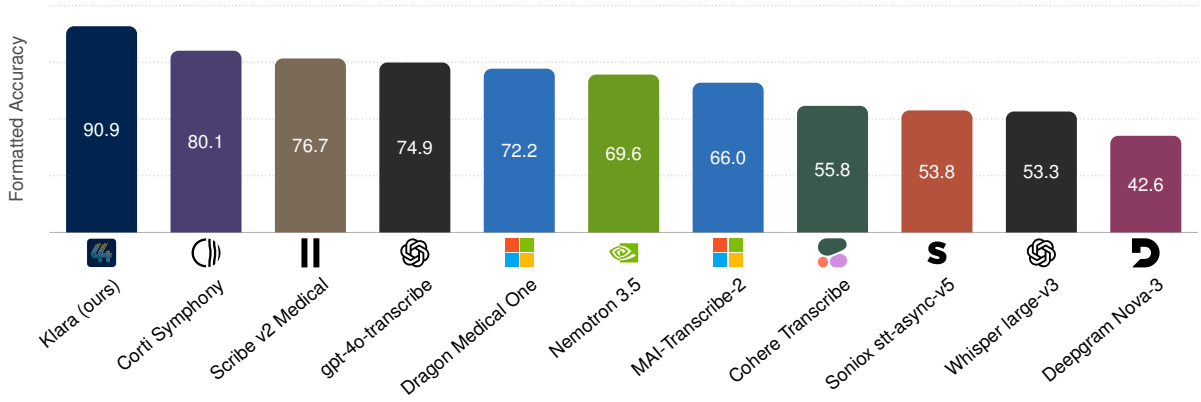
The corpus combines real clinical recordings with synthetic examples that extend coverage to rare and safety-critical terminology, unseen clinical topics, and a broad range of speaker and style variability. References follow medical documentation conventions. Audio is augmented with realistic microphone, noise, and codec variation.

2.2 Dialects

The corpus spans five groups: Swiss German dialects, German regional dialects, High German, Austrian German, and non-native German. Swiss German speech is transcribed to standard German, matching the diglossic documentation practice of Swiss clinics (Plüss et al., 2023). Subsets are balanced by target hours per group and by speaker gender, so low-resource dialects are not drowned out by standard German.

2.3 Medical Terms

The terminology database contains more than 150'000 German medical terms: medications, diagnoses, procedures, anatomy, and syndromes. Rare



#	System	Combined	per test set		
			44ai-clinical-de	MedTerm	MedDictate
1	Klara (ours)	**90.9**	**90.0**	93.5	**91.5**
2	Corti Symphony	80.1	74.5	**95.2**	87.9
3	ElevenLabs Scribe$_{\text{v2 Medical}}$	76.7	72.7	87.5	83.4
4	gpt-4o-transcribe	74.9	73.5	77.1	86.7
5	Dragon Medical One	72.2*	66.5	86.5*	87.0*
6	ElevenLabs Scribe$_{\text{v2}}$	72.1	66.9	86.6	76.4
7	Nemotron 3.5 Transcribe	69.6	65.5	80.8	73.1
8	MAI-Transcribe-2	66.0	62.1	79.0	52.2
9	Cohere Transcribe	55.8*	52.6	66.0*	48.4*
10	Soniox stt-async-v5	53.8	46.2	73.6	67.7
11	Whisper large-v3	53.3*	47.7	70.1*	46.8*
12	Deepgram Nova-3	42.6	33.9	69.3	31.0

Table 1: Systems ranked by formatted accuracy (100 − fWER, %) over all three test sets, weighted by audio duration. *Run on the chunked MedTerm and MedDictate (Section 3). Best in **bold**.

and safety-critical terms are overrepresented in the generated narratives, so the model sees them in varied, realistic contexts.

3 Test Sets

We evaluate on `44ai-clinical-de`, a clinical dictation set whose medical topics are held out from training, released on Hugging Face[2].

Composition 4'331 dictations, 12 hours of audio, covering nine dialects and accents in the five groups of Section 2.2 with balanced speaker gender: roughly 310-370 samples per dialect and 1'505 for High German. Every sample carries its reference transcript in spoken and formatted form, its dialect tag, and the medical terms it contains.

Corti benchmarks We also evaluate on the German portions of Corti's MedTerm and MedDictate benchmarks (Nix et al., 2026). MedTerm has 200 synthetic, TTS-dictated recordings (4.1 hours) and MedDictate 9 real dictations (0.5 hours), both released by Corti as full-length recordings. Three systems, Cohere Transcribe, Whisper large-v3, and Dragon Medical One, were run on our chunked version of both sets, with clips of at most 30 seconds[3]. Their results on these sets are marked with an asterisk.

4 Results

4.1 Setup

We compare Klara against eight general-purpose systems, ElevenLabs Scribe v2 (ElevenLabs, 2026a), Microsoft MAI-Transcribe-2 (Microsoft AI, 2026), OpenAI gpt-4o-transcribe (OpenAI, 2025), Soniox stt-async-v5 (Soniox, 2026), Deepgram Nova-3 (Deepgram, 2025), OpenAI Whisper large-v3 (Radford et al., 2023), Cohere Transcribe (Cohere Labs, 2026), and NVIDIA Nemotron 3.5 Transcribe (NVIDIA, 2026), and three medical systems, ElevenLabs Scribe v2 Medical (ElevenLabs, 2026b), Corti Symphony (Nix et al., 2026), and Dragon Medical One (DMO) (Microsoft, 2026).

[2] https://huggingface.co/datasets/44ai/44ai-clinical-de-validation

[3] https://huggingface.co/datasets/44ai/44ai-corti-med-term-de-chunked and https://huggingface.co/datasets/44ai/44ai-corti-med-dictate-de-chunked.

System	Swiss		German regional		Austrian		All dialects	
	Word Accuracy	Medical Accuracy	Word Accuracy	Medical Accuracy	Word Accuracy	Medical Accuracy	Word Accuracy	Medical Accuracy
Klara (ours)	**90.7**	**88.0**	**92.5**	**90.0**	**93.0**	**91.9**	**91.8**	**89.4**
gpt-4o-transcribe	81.7	81.7	87.1	86.4	92.1	90.8	85.4	85.1
ElevenLabs Scribe$_{\text{v2 Medical}}$	81.3	81.4	86.2	86.3	91.0	89.8	84.7	84.7
MAI-Transcribe-2	80.3	76.5	85.8	80.0	92.7	86.7	84.4	79.6
ElevenLabs Scribe$_{\text{v2}}$	80.0	79.5	84.4	84.2	91.7	90.5	83.5	83.2
Corti Symphony	77.0	80.6	86.0	86.7	91.4	90.8	82.8	84.7
Soniox stt-async-v5	74.9	74.9	81.5	79.0	89.9	86.3	79.8	78.4
Nemotron 3.5 Transcribe	69.8	68.8	83.6	79.0	89.3	84.3	78.4	75.5
Cohere Transcribe	70.5	67.6	76.5	72.6	83.8	79.9	74.9	71.6
Whisper large-v3	69.3	67.2	73.9	73.2	84.7	81.4	73.4	71.9
Dragon Medical One	54.1	59.9	77.5	75.3	84.3	81.8	68.3	69.8
Deepgram Nova-3	51.7	54.3	74.4	70.0	84.7	79.0	66.0	64.7

Table 2: **Dialect-only comparison** Word accuracy (100 − WER) and medical term accuracy (100 − MER) in % on the Swiss German, German regional, and Austrian speakers of `44ai-clinical-de`, pooled per group. Sorted by word accuracy over all dialects, best in **bold**. Every system is scored against the spoken-form reference.

System	Combined		`44ai-clinical-de`		MedTerm$_{\text{by Corti}}$		MedDictate$_{\text{by Corti}}$	
	Word Accuracy	Medical Accuracy	Word Accuracy	Medical Accuracy	Word Accuracy	Medical Accuracy	Word Accuracy	Medical Accuracy
Klara (ours)	**93.9**	**90.4**	**92.7**	**90.6**	97.1	89.5	94.2	93.0
gpt-4o-transcribe	90.9	88.5	88.6	87.9	96.9	89.5	**95.9**	**96.0**
ElevenLabs Scribe$_{\text{v2 Medical}}$	90.7	89.0	88.3	87.4	97.4	93.0	91.2	94.2
MAI-Transcribe-2	90.6	84.6	88.3	83.0	96.8	88.5	92.4	90.0
ElevenLabs Scribe$_{\text{v2}}$	90.0	87.8	87.3	85.9	97.1	92.3	93.2	93.8
Corti Symphony	89.9	89.6	87.0	87.6	**97.8**	**94.9**	93.0	92.5
Soniox stt-async-v5†	88.0	83.6	84.8	82.0	96.4	87.0	92.1	92.7
Nemotron 3.5 Transcribe	87.3	81.5	84.1	79.9	95.9	85.0	90.4	88.6
Cohere Transcribe†	82.5*	78.1*	79.6	75.7	90.6*	84.7*	84.2*	79.8*
Whisper large-v3†	82.1*	78.3*	78.1	76.4	93.6*	83.2*	81.8*	82.7*
Dragon Medical One	80.4*	76.7*	75.6	74.6	93.1*	80.7*	89.5*	91.4*
Deepgram Nova-3†	80.0	73.3	75.1	71.3	93.1	78.3	86.7	80.2

Table 3: Word accuracy (100 − WER) and medical term accuracy (100 − MER) in % after the shared normalization (Section 4.2): combined over all three test sets, weighted by audio duration, and per test set, on `44ai-clinical-de` and the German portions of Corti's MedTerm and MedDictate. Sorted by combined word accuracy, best in **bold**. *Run on the chunked MedTerm and MedDictate (Section 3). †On MedTerm and MedDictate scored against the spoken-form reference (Section 4.2).

Test sets are `44ai-clinical-de`, MedTerm, and MedDictate (Section 3).

Commercial systems are called through their public batch APIs (September to November 2026, Table 4) with German as the language and no custom vocabulary or keyterm biasing. Each system is evaluated in one configuration. Where a system can deliver formatted text, that configuration is used for every metric: Corti Symphony with spoken and automatic punctuation, MAI-Transcribe-2 in its *clean* style, and ElevenLabs Scribe v2, Scribe v2 Medical, gpt-4o-transcribe, and Nemotron 3.5 Transcribe with one shared formatting instruction, which ships with the harness and is passed as ElevenLabs' transcript edit instruction and as the prompt of the other two. The instruction states German dictation conventions and is the only prompt in the reported configurations. Deepgram Nova-3 runs with its smart formatting, which leaves dictated commands as words. Soniox, Whisper large-v3, and Cohere Transcribe have no formatting step. Table 5 also lists the runs without the formatting step. Whisper large-v3 and Cohere Transcribe run locally with greedy decoding, Nemotron 3.5 Transcribe with the serving setup and settings of its model card. Dragon Medical One (2026.1, *Deutsch (Schweiz)*, *Medizin - allgemein*) has no batch interface and is driven through its desktop client. Klara is called through 44ai's production stack, without a caller-supplied glossary. The exact invocation code for every system ships with our evaluation harness[4].

[4] `44ai-labs/44ai-klara-evaluation`

4.2 Scoring

All numbers are computed with the released harness, which holds the frozen transcripts of every system and regenerates every table. Its code is the full specification of the rules summarized here.

Normalization Reference, hypothesis, and medical-term list pass through the same pipeline, which lowercases and maps numbers, units, dates, symbols, and spoken punctuation to one written form, and folds compound spellings, attested Latinate variants (*occipital*/*okzipital*), common clinical abbreviations, umlauts, and *ß*. There is no blanket character rewriting, so mis-hearings such as *Torax* for *Thorax* remain errors (Manohar and Pillai, 2024).

Metrics *Word accuracy* is $100 - \text{WER}$, where WER is the word error rate over the normalized text. Punctuation a system adds is not charged. On `44ai-clinical-de`, every system is scored against the spoken-form reference. On MedTerm and MedDictate, the systems that deliver formatted text are scored against the formatted reference. *Medical term accuracy* is $100 - \text{MER}$, where MER, the medical term error rate, is one minus the recall of the reference's medical terms (Afonja et al., 2024; Corti, 2026), using each test set's own term annotations. A term split into or glued to other words still counts as recognized (Uzuner et al., 2011; Segura-Bedmar et al., 2013). *Formatted WER (fWER)* scores the raw output against the formatted reference without normalization, so capitalization, punctuation, and formatting count. For systems run on the chunked Corti sets, the clip transcripts are joined per recording before scoring. Combined scores weight each test set by its audio duration.

4.3 Main Results

Klara leads the combined ranking on formatted accuracy (90.9%, Table 1) and on medical term accuracy (90.4%, Table 3). On `44ai-clinical-de`, it has the highest word accuracy (92.7% vs. 88.6% for gpt-4o-transcribe, 36% fewer word errors) and the highest medical term accuracy (90.6% vs. 87.9%, Table 3).

Dialects The margin is largest on dialectal speech (Table 2). Pooled over the seven Swiss German, German regional, and Austrian dialects, Klara reaches 91.8% word accuracy and 89.4% medical term accuracy against 85.4% and 85.1% for gpt-4o-transcribe. On Swiss German, Klara makes about half the word errors of the best competitor (90.7% vs. 81.7%), and it has the highest word and medical term accuracy in every dialect group.

MedTerm and MedDictate On Corti's standard-German benchmarks, Klara is within two points of the best word accuracy among full-length runs, fourth on MedTerm (97.1% vs. 97.8% for Corti Symphony) and second on MedDictate (94.2% vs. 95.9% for gpt-4o-transcribe, Table 3), and has the highest formatted accuracy on MedDictate (Table 1).

Formatted output Clinicians receive formatted text. Klara's correction stage delivers capitalized, punctuated text with formatted dosages and dates, and has the highest formatted accuracy on `44ai-clinical-de` (90.0%) and MedDictate (91.5%, Table 1). The chart above Table 1 shows the stronger of the two ElevenLabs models. Table 5 shows how each system treats dictated commands and numbers.

5 Conclusion

Klara combines a strong multilingual speech recognition backbone with a training corpus built around the failure modes of general-purpose systems: dialectal speech, dense medical terminology, and clinic acoustics. On a 12-hour multi-dialect clinical validation set, it has the highest word accuracy and medical term accuracy of twelve systems, with the largest margin on dialectal speech, and remains competitive on third-party benchmarks. The validation set and the scoring code are public, and we invite other systems to run the same evaluation.

References

Ayo Adedeji, Sarita Joshi, and Brendan Doohan. 2024. The sound of healthcare: Improving medical transcription ASR accuracy with large language models. *arXiv preprint arXiv:2402.07658*.

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant ASR models for medical entities in accented speech. In *Proceedings of Interspeech 2024*, pages 2315–2319.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation. In *Interspeech 2025*, pages 913–917. International Speech Communication Association.

Cohere Labs. 2026. Introducing cohere transcribe: State-of-the-art speech recognition. https://huggingface.co/CohereLabs/cohere-transcribe-03-2026. 2B-parameter Conformer encoder-decoder ASR model, Apache 2.0.

Corti. 2026. How to evaluate automatic speech recognition. https://www.corti.ai/guides/how-to-evaluate-automatic-speech-recognition.

Deepgram. 2025. Introducing Nova-3: Setting a new standard for AI-driven speech-to-text. https://deepgram.com/learn/introducing-nova-3-speech-to-text-api.

Eyal Dolev, Clemens Lutz, and Noëmi Aepli. 2024. Does Whisper understand Swiss German? an automatic, qualitative, and human evaluation. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 28–40, Mexico City, Mexico. Association for Computational Linguistics.

ElevenLabs. 2026a. Introducing scribe v2. https://elevenlabs.io/blog/introducing-scribe-v2.

ElevenLabs. 2026b. Scribe v2 medical is now available to everyone. https://elevenlabs.io/blog/scribe-v2-medical-generally-available.

Foster R. Goss, Suzanne V. Blackley, Carlos A. Ortega, Leigh T. Kowalski, Adam B. Landman, Chen-Tan Lin, Marie Meteer, Samantha Bakes, Stephen C. Gradwohl, David W. Bates, and Li Zhou. 2019. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *International Journal of Medical Informatics*, 130:103938.

Erica Ha, Isabelle Choon-Kon-Yune, Lucy Murray, Shirley Luan, Enid Montague, Onil Bhattacharyya, and Payal Agarwal. 2025. Evaluating the usability, technical performance, and accuracy of artificial intelligence scribes for primary care: Competitive analysis. *JMIR Human Factors*, 12:e71434.

Tobias Hodgson, Farah Magrabi, and Enrico Coiera. 2017. Efficiency and safety of speech recognition for documentation in the electronic health record. *Journal of the American Medical Informatics Association*, 24(6):1127–1133.

Kavya Manohar and Leena G. Pillai. 2024. What is lost in normalization? Exploring pitfalls in multilingual ASR model evaluations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10864–10869, Miami, Florida, USA. Association for Computational Linguistics.

Microsoft. 2026. Dragon medical one. https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-medical-one.

Microsoft AI. 2026. MAI-Transcribe-2. https://microsoft.ai/models/mai-transcribe-2/.

Arne Nix, Robert James, Lasse Borgholt, Anna B. Ekner, Lana Krumm, Julius Severin, Dan Engel, Lars Maaløe, and Jakob Havtorn. 2026. Symphony for speech-to-text: Supporting real-time medical voice interfaces. *arXiv preprint arXiv:2605.16545*.

NVIDIA. 2026. NVIDIA Nemotron 3.5 Transcribe. https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Transcribe-30B-A4B. Open-weights speech language model.

OpenAI. 2025. Introducing our next-generation audio models in the api. https://openai.com/index/introducing-our-next-generation-audio-models/. Gpt-4o-transcribe and gpt-4o-mini-transcribe.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Juan C Quiroz, Liliana Laranjo, Ahmet Baki Kocaballi, Shlomo Berkovsky, Dana Rezazadegan, and Enrico Coiera. 2019. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ digital medicine*, 2(1):114.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Thomas Schuster, J. Trögele, N. Döring, R. Krüger, M. Hoffmann, and H. Friedrich. 2026. Benchmarking ASR models in German medical contexts: A performance analysis using anamnesis conversations. In *Business Information Systems (BIS 2026)*, volume 584 of *Lecture Notes in Business Information Processing*. Springer.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Clément Sicard, Victor Gillioz, and Kajetan Pyszkowski. 2023. Spaiche: Extending state-of-the-art ASR models to Swiss German dialects. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 76–83.

Soniox. 2026. Soniox speech-to-text models. https://soniox.com/docs/stt/models.

Vincenzo Timmel, Claudio Paonessa, Manfred Vogel, Daniel Perruchoud, and Reza Kakooee. 2025. Fine-tuning Whisper on low-resource languages for real-world applications. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 57–65, Winterthur, Switzerland. Association for Computational Linguistics.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Marieke Meija van Buchem, Ilse MJ Kant, Liza King, Jacqueline Kazmaier, Ewout W Steyerberg, and Martijn P Bauer. 2024. Impact of a digital scribe system on clinical documentation time and quality: usability study. *Jmir ai*, 3(1):e60020.

Li Zhou, Suzanne V. Blackley, Leigh Kowalski, Raymond Doan, Warren W. Acker, Adam B. Landman, Evgeni Kontrient, David Mack, Marie Meteer, David W. Bates, and Foster R. Goss. 2018. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA Network Open*, 1(3):e180530.

System	Evaluated
Corti Symphony	24 Sep 2026
Deepgram Nova-3	24 Sep 2026
Soniox stt-async-v5	24 Sep 2026
ElevenLabs Scribe$_{\text{v2 Medical}}$	17 Nov 2026
ElevenLabs Scribe$_{\text{v2}}$	17 Nov 2026
gpt-4o-transcribe	17 Nov 2026
MAI-Transcribe-2	17 Nov 2026

Table 4: **API evaluation dates** Vendor API calls come with no control over the deployed model or its configuration. The dates show when we last evaluated each system.

System	Formatting	Formatted Accuracy	Word Accuracy	Medical Accuracy	Formatted Commands	Formatted Numbers
Klara (ours)	built in	**90.9**	**93.9**	**90.4**	**100**	**100**
Corti Symphony	formatting options[2]	80.1	89.9	89.6	97	88
ElevenLabs Scribe$_{\text{v2 Medical}}$	shared instruction[1]	76.7	90.7	89.0	85	74
gpt-4o-transcribe	shared instruction[1]	74.9	90.9	88.5	80	90
Dragon Medical One	built in	72.2*	80.4*	76.7*	**100**	**100**
ElevenLabs Scribe$_{\text{v2}}$	shared instruction[1]	72.1	90.0	87.8	82	44
gpt-4o-transcribe	API default	70.0	87.9	86.4	76	85
Nemotron 3.5 Transcribe	shared instruction[1]	69.6	87.3	81.5	74	82
ElevenLabs Scribe$_{\text{v2 Medical}}$	API default	68.8	92.4	89.2	28	66
MAI-Transcribe-2	clean style[3]	66.0	90.6	84.6	53	93
Corti Symphony	as dictated[2]	64.8	90.3	89.7	2	86
ElevenLabs Scribe$_{\text{v2}}$	API default	59.5	91.6	88.4	20	30
MAI-Transcribe-2	verbatim style[3]	59.0	90.5	84.7	30	96
Nemotron 3.5 Transcribe	model default	57.7	85.1	79.6	53	12
Cohere Transcribe	model default	55.8*	82.5*	78.1*	57	67
Soniox stt-async-v5	API default	53.8	88.0	83.6	20	**100**
Whisper large-v3	model default	53.3*	82.1*	78.3*	45	91
Deepgram Nova-3	smart format[4]	42.6	80.0	73.3	5	96

[1] github.com/44ai-labs/44ai-klara-evaluation/blob/main/src/klara_benchmark/systems/formatting.py#L3
[2] docs.corti.ai/stt/formatting
[3] learn.microsoft.com/azure/ai-services/speech-service/mai-transcribe
[4] developers.deepgram.com/docs/smart-format

Table 5: **All configurations** Formatted, word, and medical term accuracy in % over all three test sets, weighted by audio duration, sorted by formatted accuracy. *Formatted commands*: share of the `44ai-clinical-de` clips with a dictated command (*Punkt*, *Komma*, *Neue Zeile*, . . .) in which no command is left as a word. *Formatted numbers*: share of the clips with numbers in which none is written out. *Run on the chunked MedTerm and MedDictate (Section 3).