Berta: An open-source, modular tool for AI-enabled clinical documentation

Samridhi Vaid[1], **Mike Weldon**[2,3], **Jesse Dunn**[3], **Kevin Lonergan**[3], **Henry Li**[2,3], **Jeffrey Franc**[2,3], **Mohamed Abdala**[1,4,5], **Daniel C. Baumgart**[1], **Jake Hayward**[2,3], and **J Ross Mitchell**[1,4,5¶]

1 Department of Medicine, University of Alberta, Edmonton, Alberta, Canada **2** Department of Emergency Medicine, University of Alberta, Edmonton, Alberta, Canada **3** Alberta Health Services, Edmonton, Alberta, Canada **4** Department of Computer Science, University of Alberta, Edmonton, Alberta, Canada **5** Alberta Machine Intelligence Institute, Edmonton, Alberta, Canada ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Berta is an open-source, modular platform for building and evaluating AI-enabled clinical documentation systems. Named in homage to Alberta and BERT (Bidirectional Encoder Representations from Transformers), Berta combines automatic speech recognition (ASR) with large language models (LLMs) to transcribe patient encounters and generate structured clinical notes. The system comprises a Python FastAPI (Ramirez, 2018) backend and a Next.js frontend, and supports deployment on systems ranging from a single workstation to a GPU server in a secure virtual private cloud, to cloud environments such as Amazon Web Services.

The Journal of Open Source Software

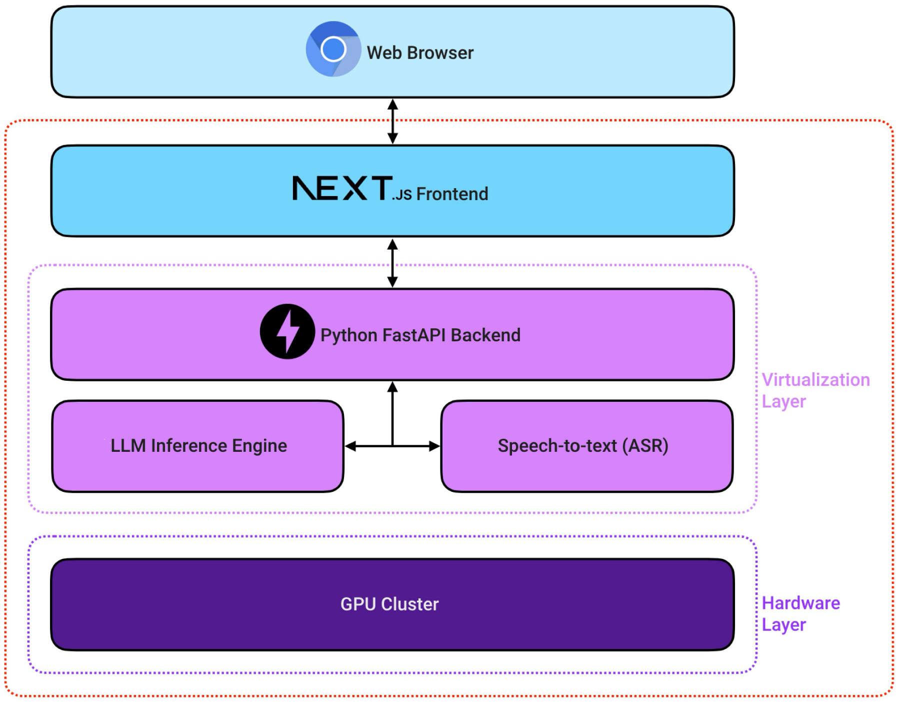

Figure 1: Berta system architecture. The system features a Next.js frontend, a Python FastAPI backend, and modular ASR and LLM components that can be deployed on-premises or in a virtual private cloud.

Statement of need

Emergency physicians in developed countries typically spend more than 40% of their time on documentation and less than 30% on direct patient care (Hill et al., 2013). This administrative burden is a major contributor to physician burnout (Shanafelt et al., 2016), reduced career satisfaction (Melnick et al., 2021), and workforce attrition. The financial impact is substantial, with burnout-related physician turnover costing an estimated US$4.6 billion annually in the United States (Han et al., 2019), while emergency medicine reports burnout rates of up to 86% among physicians in developed countries (Lim et al., 2023). The consequences extend throughout healthcare systems: Canada experienced more than 1,200 temporary emergency department closures in 2023 alone, disproportionately affecting rural and underserved communities (CTV News, 2023).

Electronic transcription solutions (scribes) can reduce documentation time by up to 35% (Hess et al., 2015) and increase patient throughput by 10–20% (Walker et al., 2019). However, current commercial AI scribe solutions often operate as expensive proprietary "black-box" systems with limited transparency (Kim et al., 2025), costing several hundred dollars per physician per month (Heidi Health, 2025; Scribeberry, 2025) and restricting organizational control over data governance and system customization (NHS England, 2025). Healthcare organizations, particularly those in resource-constrained environments, lack accessible tools to evaluate, customize, and deploy AI documentation systems according to their specific clinical workflows and regulatory requirements (Wong et al., 2025).

Berta addresses this gap by providing an open-source modular platform that enables healthcare organizations to build, test, and deploy AI-powered clinical documentation systems with full transparency, data sovereignty, and cost-effective scalability, supporting informed decision-making about this rapidly evolving technology.

State of the field

Commercial AI scribe products are closed-source, subscription-based services with vendor-reported estimates ranging from US$99 to over US$600 per physician per month (Heidi Health, 2025; Scribeberry, 2025). These systems offer polished integrations with electronic health records but provide no access to source code, limit customization of note templates and model selection, and require organizations to route clinical audio through third-party infrastructure. Their proprietary nature makes independent auditing, bias evaluation, and regulatory compliance verification difficult (Kim et al., 2025).

Berta is not intended to compete with commercial AI scribe products. Rather, it is intended to help organizations evaluate AI-enabled clinical documentation systems and gather information to guide future decision-making. To our knowledge, no comparable open-source tool exists that provides a complete, deployment-ready platform for AI-enabled clinical documentation with modular ASR and LLM backends.

Software design

Berta comprises a Next.js frontend and a FastAPI (Ramirez, 2018) backend that exposes RESTful APIs for application logic, data processing, and system integration (Figure 1). In routine use, clinicians create a session in the web application and record or upload audio from a patient encounter. The system transcribes speech with an ASR model and then uses an LLM to generate a structured draft clinical note from the transcript using configurable note templates (e.g., full visit note, narrative, handover summary); users can also create and save custom templates. Clinicians review and edit the generated note before transferring it to their electronic health record.

The platform adopts a modular adapter pattern across its ASR and LLM components. Supported ASR backends include WhisperX (Bain et al., 2023), OpenAI Whisper (Radford et al., 2023), NVIDIA Parakeet via MLX (Hannun et al., 2023; NVIDIA, 2025; senstella, 2025), and Amazon Transcribe (Amazon Web Services, Inc., 2025b); supported LLM backends include local engines (Ollama (Ollama, 2023), vLLM (Kwon & others, 2023), LM Studio (LM Studio, 2024)) and commercial endpoints (OpenAI API (OpenAI, 2025), Amazon Bedrock (Amazon Web Services, Inc., 2025a)). This modular design allows organizations to interchange backends without modifying application code. Clinicians can customize note templates and prompts to match their charting preferences, and all data can be retained on-premises or within a chosen cloud environment, giving organizations full control over data sovereignty.

Research impact statement

A closed-source deployment of the platform underlying Berta has been operational at Alberta Health Services (AHS) since November 2024. During the pilot period, the system was used in 22,148 sessions by 198 emergency physicians across 105 healthcare facilities in Alberta, Canada, representing a mix of urban and rural settings. Approximately 42% of users customized at least one document template to align with their individual charting preferences. Based on observed usage, the average operating cost of delivering the application was less than US$30 per physician per month, demonstrating that high-volume provincial-scale clinical use can be sustained at relatively low per-physician cost. Based on pilot results, AHS expanded access and has since invited over 1,600 emergency department physicians.

AI usage disclosure

Claude Code (Anthropic Claude Opus 4.5 and 4.6) and ChatGPT (OpenAI GPT-4o) were used for code assistance, debugging, and test generation during development. All AI-generated code was reviewed and validated by the development team. Claude (Anthropic) was used to assist with structuring and reviewing drafts of this paper. The final text was written and verified by the authors.

Acknowledgements

This work was supported by the Canadian Medical Association, MD Financial Management, and Scotiabank through the Health Care Unburdened Grant program. We acknowledge the support provided by the Canadian Institute for Advanced Research, the University Hospital Foundation, Alberta Health Services, Amazon Web Services, and Denvr Dataworks. This project uses third-party libraries and models, including WhisperX (BSD 2-Clause), Meta Llama 3 (Meta Llama 3 Community License), NVIDIA Parakeet (CC-BY-4.0), vLLM (Apache 2.0), and Ollama (MIT License).

References

Amazon Web Services, Inc. (2025a). *Amazon bedrock*. https://aws.amazon.com/bedrock/.

Amazon Web Services, Inc. (2025b). *Amazon transcribe*. https://aws.amazon.com/transcribe/.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. *arXiv Preprint arXiv:2303.00747*. https://doi.org/10.48550/arXiv.2303.00747

CTV News. (2023). Three stabbed teens were driven from a party to a nearby hospital, only to find that the ER was closed: Their story is one of many. *CTV News*. https://www.ctvnews.ca/health/three-stabbed-teens-were-driven-from-a-party-to-a-nearby-hospital-only-to-find-that-the-er-was-closed-their-story-is-one-of-many-1.6545043

Han, S., Shanafelt, T. D., Sinsky, C. A., Awad, K. M., Dyrbye, L. N., Fiscus, L. C., Trockel, M., & Goh, J. (2019). Estimating the attributable cost of physician burnout in the United States. *Annals of Internal Medicine*, *170*(11), 784–790. https://doi.org/10.7326/M18-1422

Hannun, A., Digani, J., Katharopoulos, A., & Collobert, R. (2023). *MLX: Efficient and flexible machine learning on Apple silicon*. https://github.com/ml-explore/mlx.

Heidi Health. (2025). *AI medical scribe cost: Is it worth the price?* https://www.heidihealth.com/en-ca/blog/ai-medical-scribe-cost.

Hess, J. J., Wallenstein, J., Ackerman, J. D., Akhter, M., Ander, D., Keadey, M. T., & Capes, J. P. (2015). Scribe impacts on provider experience, operations, and teaching in an academic emergency medicine practice. *Western Journal of Emergency Medicine*, *16*, 602–610. https://doi.org/10.5811/westjem.2015.6.25082

Hill, R. G., Sears, L. M., & Melanson, S. W. (2013). 4000 clicks: A productivity analysis of electronic medical records in a community hospital ED. *American Journal of Emergency Medicine*, *31*, 1591–1594. https://doi.org/10.1016/j.ajem.2013.06.028

Kim, C., Gadgil, S. U., & Lee, S.-I. (2025). Transparency of medical artificial intelligence systems. *Nature Reviews Bioengineering*, 1–19. https://doi.org/10.1038/s44222-025-00291-9

Kwon, W., & others. (2023). Efficient memory management for large language model serving with PagedAttention. *arXiv Preprint arXiv:2309.06180*. https://doi.org/10.48550/arXiv.2309.06180

Lim, R., Aarsen, K. V., Gray, S., Rang, L., Fitzpatrick, J., & Fischer, L. (2023). Emergency medicine physician burnout and wellness in Canada before COVID-19: A national survey.

Canadian Journal of Emergency Medicine, 1–5. https://doi.org/10.1007/s43678-023-00484-y

LM Studio. (2024). *LM Studio: Discover, download, and run local LLMs*. https://lmstudio.ai/.

Melnick, E. R., Fong, A., Nath, B., Williams, B., Ratwani, R. M., Goldstein, R., O'Connell, R. T., Sinsky, C. A., Marchalik, D., & Mete, M. (2021). Analysis of electronic health record use and clinical productivity and their association with physician turnover. *JAMA Network Open*, *4*, e2128790. https://doi.org/10.1001/jamanetworkopen.2021.28790

NHS England. (2025). *Guidance on the use of AI-enabled ambient scribing products in health and care settings*. https://www.england.nhs.uk/long-read/guidance-on-the-use-of-ai-enabled-ambient-scribing-products-in-health-and-care-settings/.

NVIDIA. (2025). *Parakeet TDT 0.6B V2 (en)*. https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2.

Ollama. (2023). *Ollama: Get up and running with large language models locally*. https://github.com/ollama/ollama.

OpenAI. (2025). *OpenAI API platform*. https://platform.openai.com/.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning*, 28492–28518. https://doi.org/10.48550/arXiv.2212.04356

Ramirez, S. (2018). *FastAPI: Modern, fast web framework for building APIs with Python*. https://fastapi.tiangolo.com/.

Scribeberry. (2025). *AI vs traditional medical scribing: A cost comparison*. https://blog.scribeberry.com/ai-vs-traditional-medical-scribing-a-cost-comparison/.

senstella. (2025). *Parakeet-mlx: Parakeet speech models implemented in MLX*. https://github.com/senstella/parakeet-mlx.

Shanafelt, T. D., Dyrbye, L. N., Sinsky, C., Hasan, O., Satele, D., Sloan, J., & West, C. P. (2016). Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic Proceedings*, *91*, 836–848. https://doi.org/10.1016/j.mayocp.2016.05.007

Walker, K., Ben-Meir, M., Dunlop, W., & others. (2019). Impact of scribes on emergency medicine doctors' productivity and patient throughput: Multicentre randomised trial. *BMJ*, *364*, l121. https://doi.org/10.1136/bmj.l121

Wong, E., Bermudez-Cañete, A., Campbell, M. J., & Rhew, D. C. (2025). Bridging the digital divide: A practical roadmap for deploying medical artificial intelligence technologies in low-resource settings. *Population Health Management*, *28*(2), 105–114. https://doi.org/10.1089/pop.2024.0250