University Hospital Bonn on-premise language model

IMPLEMENTED 2025–2026 · HEALTH AND LIFE SCIENCES · GERMANY · BAND D · CLAIMED ONLY
VERSION 1.0 · PUBLISHED 15 NOVEMBER 2026 · UPDATED 15 NOVEMBER 2026 · FACTS CHECKED 15 NOVEMBER 2026

University Hospital Bonn runs the open-weight DeepSeek-R1 model on a server in its own data centre so that radiologists can use a language model on patient text without it leaving the hospital. In a one-week pilot in July 2025, 22 radiologists rated it stable and most useful for tasks anchored in existing text, such as translating and correcting reports.

ABOUT THIS SYNOPSIS

This is MarLuc's summary and grading of publicly available sources, for information only. It is not investment, legal, financial or professional advice, and should not be relied on for any decision. Readers must validate it against the original sources and, where appropriate, challenge its conclusions.

Facts and figures are as reported by the sources listed below, which may be incomplete, self-reported or out of date. Check the original source before relying on any figure.

The impact band, proof marker and SEAM view are MarLuc's opinion. The organisation is not a MarLuc client and has not reviewed or endorsed this synopsis. Names and trademarks belong to their owners.

NOTICE WORDING LAST UPDATED 16 NOVEMBER 2026

AT A GLANCE

AI used: Generative AI assistant: open-weight DeepSeek-R1 model hosted on the hospital's own server.

Implemented: 2025–2026. Grade: band D, Task support, Claimed only.

Headline result, as reported: Open-weight DeepSeek-R1 run inside the hospital data centre; 22 radiologists rated it stable in a one-week pilot, by the hospital's own account

1 · The case

The facts, as reported by the 3 sources listed at the end. Where a source is silent, the field says "Not reported".

Organisation

University Hospital Bonn. University hospital in Bonn, Germany. The authors describe it as a university hospital with over 10,000 employees. Canvas block: 6. Data and technology. Also 8. Governance and risk, 4. Operations and process.

Business challenge

The challenge. The authors say that well-known commercial models are offered through cloud interfaces that need data to leave the hospital network, which raises confidentiality and regulatory concerns for patient information. Their aim was a self-hosted service that could

Who implemented it

DeepSeek (Model or cloud provider): Provides the DeepSeek-R1 open-weight model used in the pilot, as named in the paper..
In-house (In-house): University Hospital Bonn Research-IT and Department of Diagnostic and Interventional Radiology: Designed, built, hosted and evaluated the system on the hospital's own server..
The authors released the deployment package under the Apache-2.0 licence on GitHub. No consultancy or integrator is named.

Technology (SEAM pillar E)

SEAM types: T6 Open-source / locally hosted models. Deployment: Own infrastructure. Vendors and products: DeepSeek, DeepSeek-R1 (671 billion parameters, quantised), vLLM, Open WebUI, UKB-GPT (open-source deployment package).

Where it was built in: A chat interface for radiology text tasks on a single server with eight NVIDIA A100 80 GB GPUs in the hospital data centre, reached from the hospital intranet, with staff sign-in through the hospital's directory service.

What changed

Before: the authors say hospitals in many countries do not permit external language model services with clinical data. After: staff can use a locally hosted model on the hospital network, after completing an AI literacy course. [1] [2] [3]

What it cost

Not reported

What it saved

No savings reported. In the pilot, users rated ICD-10 coding as having high potential but low time savings (median 3.5), because checking the codes offered little gain over manual lookup. The authors say open models could cut licence costs, but report no figure. [1] [2] [3]

Governance and control

By the authors' account, the board, data protection officer, information security officer, works council and compliance department each set conditions: a 30-day chat retention limit, no staff usage statistics and voluntary use, a mandatory AI literacy course and a live security audit. All of this comes from the hospital's own paper.

S2 · Accountable owner · Claimed
The hospital's board of directors approved funding and the formal basis for deployment, by the authors' account; no named individual owner is reported. [1]

A1 · Safe use · Claimed
Staff must complete an AI literacy course before access, chat retention is capped at 30 days and the system is defined as radiology text support without autonomous clinical decisions, by the authors' account. [1]

A2 · Human in the loop · Claimed
The model is described as an assistive tool; users in the pilot reported where outputs needed correction, and the authors say ICD-10 coding still needed human checking. [1]

A3 · Monitoring and assurance · Claimed
Automated isolation tests and active isolation monitoring ran, and the authors report no external connectivity events in the pilot; error reports came from users, not blinded reviewers. [1]

A4 · Vendor AI · Claimed
The authors report a risk assessment of the open-source software components and defined supply chain and malicious model weights as threat scenarios. [1]

A5 · Agent controls · Not reported
The service is a chat assistant. The authors list agent workflows as future work; no agent controls are reported. [1]

A6 · Standards and regulation · Claimed
Hospital compliance concluded that the EU AI Act applies and defined the hospital as a deployer, and the data protection and information security officers and the works council approved, by the authors' account. [1]

What worked, what did not, what is unproven

WORKED

– Stability and usability, rated highly by the 22 pilot users (stability median 10.0 out of 10).

– Tasks tied to an existing text, such as translation, linguistic correction and report simplification, with high ratings and no reported critical errors.

– Isolation tests passed, and the authors report no external connectivity events during the pilot.

DID NOT

– Open-ended conclusion drafting from findings had the highest number of critical errors, including two users reporting multiple critical errors.

– ICD-10 coding saved little time because the codes had to be verified.

– The authors say the older GPU server limited speed, with about 25 output tokens per second.

UNPROVEN

– Whether the benefits hold in routine use beyond a one-week pilot with 22 radiologists.

– Whether clinical error rates are acceptable, as errors were reported by users and not reviewed blind.

– Any effect on cost, report turnaround or patient outcomes.

Timeline

Jan 2025 University Hospital Bonn and the University of Bonn report that local open-weight models can structure radiology findings with all data staying at the hospital [3]

Jul 2025 One-week pilot with 22 radiologists using ten prompt templates, after ethics approval [1]

Mar 2026 Hospital authors post a preprint describing the system as a permanent service for the hospital's staff [1]

2 · Grade and SEAM assessment

MarLuc's opinion, made on the evidence above. How cases are graded: marluc.co.uk/case-studies/how-we-rate

Level of change and impact

Claimed: Not applicable. Grade: D, Task support. Proof: Claimed only. Level of change: Task support.

Task support (Level 1, band D). Individuals use the model for text tasks; the report shows no change to a process, team structure or budget.

Radiologists use the tool for individual text tasks, so the main effect is on task support. The authors report ratings from a small one-week pilot and no cost, time or quality measures from routine use.

The SEAM view and lessons learnt

The main contribution is the control design: a locally hosted open-weight model with tested network isolation and a documented approval route. The evidence of benefit is thin, a one-week survey of 22 users. For a larger organisation, the transferable point is to limit early use to tasks anchored in existing text and to keep checking outputs by hand.

1. Start with tasks tied to a source text.
2. Treat chat logs on a local server as data that needs retention rules.
3. Agree conditions with data protection, security and staff representatives before launch.

Evidence

Score 1 of 4. Self-reported: the results come from the hospital authors' own preprint and survey; the GitHub code is public but the outcomes are not independently checked.

Evaluation method. Self-reported company figures: a one-week pilot survey of 22 radiologists and system logs; error reports came from users, not blinded reviewers.

Sources disagree. The preprint describes over 10,000 employees at the hospital. The hospital's 2023 annual report was found in search but not read in full for this synopsis, so the staff figure is the authors' own.

Status. Ongoing: the authors describe it as the base for a permanent service for hospital staff, as of March 2026.

3 · Original sources

Read these in full. To challenge a fact or a grade, use the contact form at marluc.co.uk or the link on https://marluc.co.uk/case-studies/university-hospital-bonn-on-premise-llm

[1] arXiv (University Hospital Bonn authors): Secure On-Premise Deployment of Open-Weights Large Language Models in Radiology

[2] GitHub (University Hospital Bonn): UKB-GPT, a single-host, isolation-first GenAI stack
https://github.com/ukbonn/ukb-gpt

[3] University of Bonn: From text to structured information, securely with AI (14 January 2025)
https://www.uni-bonn.de/en/news/from-text-to-structured-information-securely-with-ai

University Hospital Bonn website: https://www.ukbonn.de

Version 1.0 · Published 15 November 2026 · Updated 15 November 2026 · Facts checked against sources 15 November 2026 · Downloaded 26 November 2026. This PDF may be out of date: the latest version, and a way to report an error or a better source, is at https://marluc.co.uk/case-studies/university-hospital-bonn-on-premise-llm

VERSION HISTORY

v1.0 15 November 2026 First published