BELLADONNA

Breast Expert-Led LLM Agent for Data Organization, Normalization & Narrative Aggregation

Dr. med. Catharina Wichmann

Ärztin in Weiterbildung für Gynäkologie und Geburtshilfe

Universitätsfrauenklinik Dresden · Kather Lab · NCT Dresden

Sächsische Krebsregister-Qualitätskonferenz Mammakarzinom · Chemnitz · April 2026

BCRF BREAST CANCER RESEARCH FOUNDATION

Warum ist das Thema jetzt klinisch relevant?

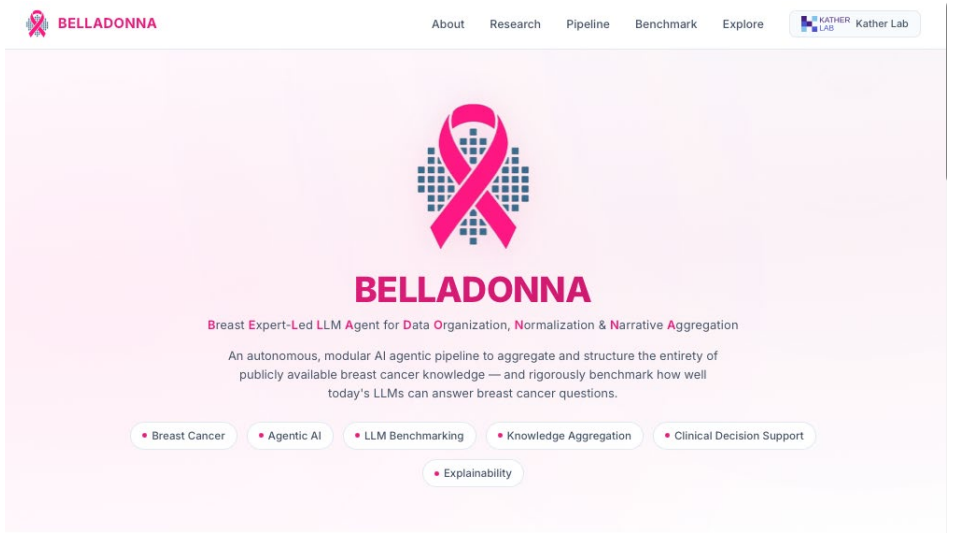
Ausgangslage

66 % Ärzt:innen nutzen KI bereits in ihrer Praxis

ca. 30% Patient:innen nutzen Large language models für Gesundheitsfragen

+ Evidenz laufend neue Leitlinien, Biomarker, Therapieoptionen

Informationsflut
kaum systematisch kuratierte Quellen



LLMs bereits Teil der klinischen Kommunikation
nicht nur für Forschung

Blease et al., BMJ Health Care Inform. 2024, American Medical Association. 2024
KFF. Poll. 2026

Was versteht man unter LLM?

Large language model = Sprachmodell

TEXT

Trainingsdaten

- sehr große Textmengen
- Paper, Leitlinien, Webseiten
- lernt Sprachmuster

PRED

Modellprinzip

- berechnet wahrscheinliche Wortfolge
- keine eigene Wahrheitsprüfung
- kein implizites klinisches Verständnis

OUT

Antwort

- flüssig, plausibel, schnell
- sprachlich oft überzeugend
- inhaltlich nicht automatisch korrekt

plausibel formuliert ≠ medizinisch valide

Singhal et al., Nature Medicine 2023

Wo liegt das Risiko in der Onkologie?

Warum BELLADONNA notwendig ist

Vorteile von LLMs

- komplexe Inhalte verdichten
- große Textmengen schnell zusammenfassen
- verständlich formulieren
- strukturierte Ausgaben erzeugen

Kritisch im klinischen Setting

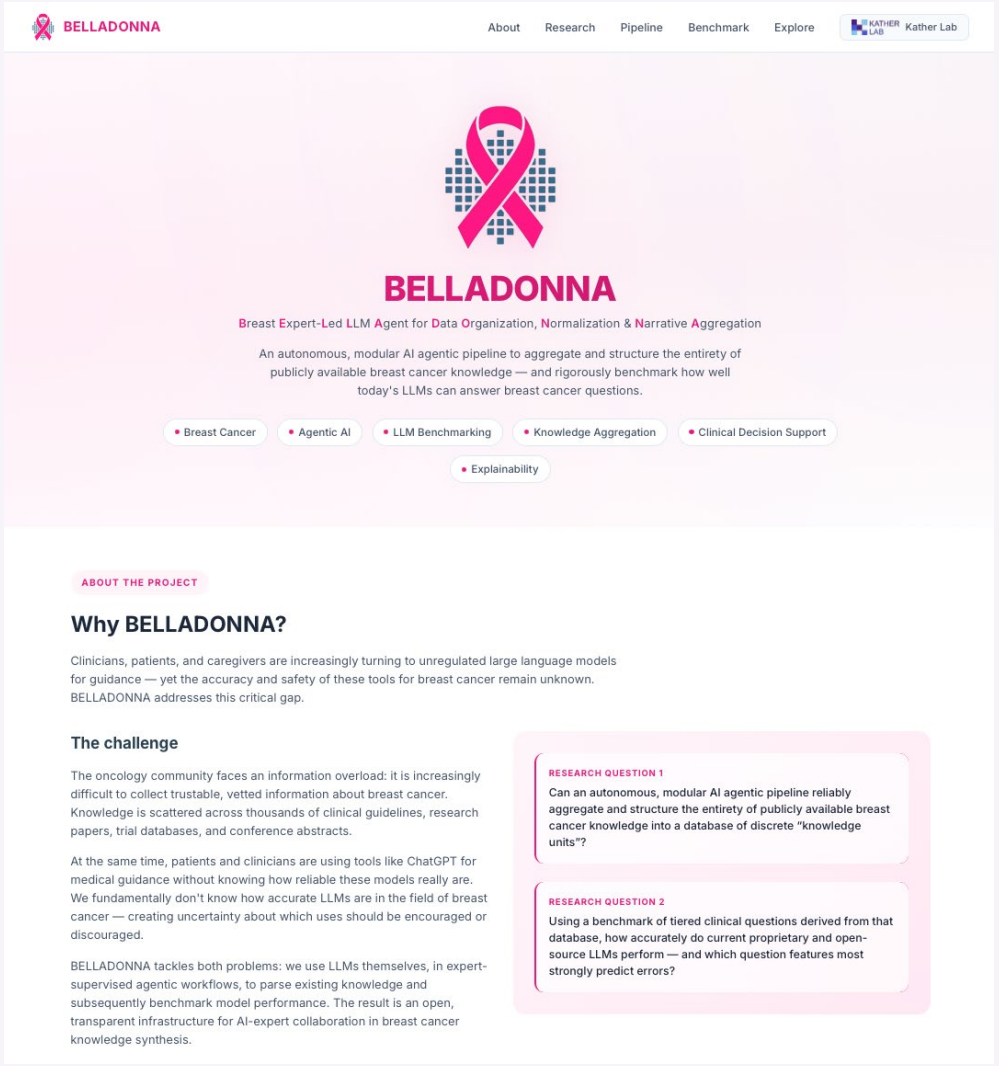
- Halluzinationen
  (= falsche Information, die fachlich plausibel klingt)
- veraltete Leitlinieninhalte
- fehlende domänenspezifische Qualitätsprüfung
- Fehler oft schwer erkennbar

→ BELLADONNA adressiert genau diese Lücke
strukturierte Wissensbasis + kontrollierte Evaluation

Lyu et al., JCO 2025; Bitterman et al., JCO 2024; Marra et al., Ann Oncol. 2025

BELLADONNA

Eine Wissensdatenbank – mehrere Anwendungen



Strukturierte Wissensbasis

- Literatur & Leitlinien priorisieren
- Wissen strukturieren
- Factoids / Datenbank aufbauen

Von Literatur zu strukturiertem Wissen

Wissensaggregations-Pipeline

1

Quellen

AGO Mamma
ESMO
EPMC · Elsevier
FDA/EMA · Trials

2

Filterung

Relevanz
Deduplizierung
Quellenpriorisierung

3

Strukturierung

Wissenseinheiten / Factoids
einheitliche Klassifikation
Domänen-Annotation

4

Anwendungen

Benchmarking
Visualisierung
Patientinnen-informationswebsite

Ziel: kuratierbare, domänenspezifische Factoids

Aktueller Projektstand

Datenbasis und Priorisierung

EPMC

354.225 Downloads
117.602 gefiltert

Elsevier

345.286 Downloads
51.206 gefiltert

weitere Literaturquellen

AGO Mamma 2025, S3 Leitlinie,
ESMO, ASCO, ClinicalTrials, FDA/EMA

Zielgröße

~ 50.000
strukturierte Factoids

Priorisierung der Factoids

1. Leitlinien
2. Literatur
3. Empfehlungen

priorisierte und nachvollziehbare Wissensdatenbank

Vom Forschungsprojekt zur Anwendung

Wofür BELLADONNA genutzt werden kann

1 **Wissensdatenbank**
strukturierte, evidenzbasierte Factoids

2 **Benchmark**
systematische Evaluation aktueller LLMs

3 **Weiterbildung**
Werkzeug für Senologie und Expert:innenbereich (Interface in Entwicklung)

4 **Patientinneninformation mit Chatbot**
vereinfachte, evidenzbasierte Sprache (Interface in Entwicklung)

BELLADONNA-website = Datenbank + Benchmark + Translation

Hoeper et al., JCO 2025; Ferber et al., NEJM AI 2024;
Lee et al., Nat Cancer 2024; GeneSilico Copilot, medRxiv 2024

Benchmarking: Human Expert vs. LLM vs. BELLADONNA

Aktuell laufendes / geplantes Studiendesign

Der Kern

200 Expert-Level Multiple choice questions

12 klinische Domänen
100 % human-authored

Human Expert Arm

- BIBD-Design
- 40 Fragen pro Person

LLM Arm

- Alle 200 Fragen pro Modell
- Standardisierte Prompts
- Mehrere Durchläufe pro Modell

BELLADONNA Knowledge Arm

- dieselben 200 Fragen
- Prüfung der Wissensbasis
- Vergleich mit Human + LLMS

Überprüfung von Accuracy, Fehlermuster der LLMs & Mehrwert der strukturierten Datenbank

Singhal et al., Nature Medicine 2023;
Ferber et al., arXiv 2024; Ferber et al., NEJM AI 2024

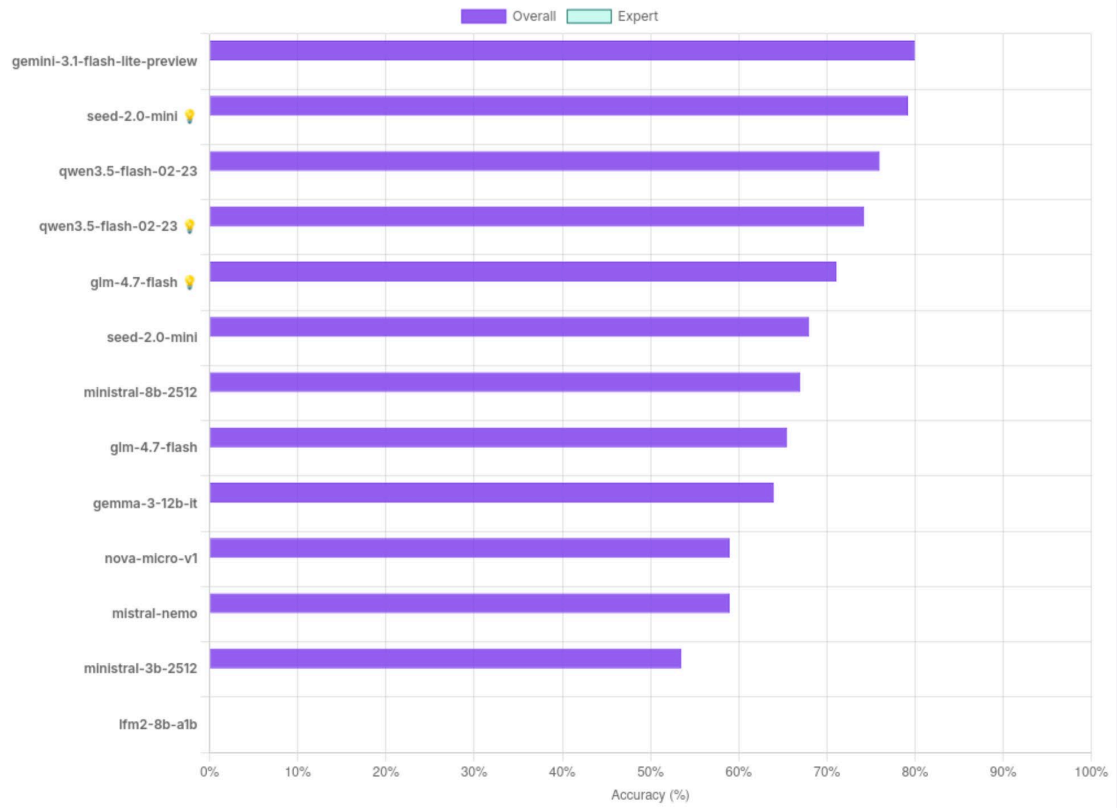
Pre-set MCQ-Ergebnisse: light weight LLM-Arm

12 Sprachmodelle getestet – Human Expert Arm ausstehend

Benchmark Results

Model performance across the available benchmark results. Last updated: 3/27/2026, 2:28:20 PM

Overall Comparison



Accuracy Heatmap

MODEL	OVERALL	EXPERT...	BIOMAR...	CURATIVE	EPIDEMI...	GENETICS	METAST...	SCREENI...	SPECIAL...	SUPPOR...	SURGERY	SYSTEMI...	TOXICITY	TRIAL
gemini-3.1-flash-lite-pr...	80.0%	80	83	74	83	100	63	84	67	80	87	82	82	90
seed-2.0-mini	79.2%	79	50	68	78	100	48	89	44	60	87	64	82	70
qwen3.5-flash-02-23...	76.0%	76	50	66	83	100	54	100	67	80	80	73	82	80
qwen3.5-flash-02-23 ...	74.2%	74	83	74	78	100	46	79	44	80	87	64	91	70
glm-4.7-flash 30B/3B	71.1%	71	40	71	76	100	53	79	43	70	83	60	80	67
seed-2.0-mini	68.0%	68	50	61	83	94	50	84	33	50	73	59	82	90
ministral-8b-2512 8B	67.0%	67	33	66	83	94	38	79	44	60	87	59	64	80
glm-4.7-flash 30B/3B	65.5%	66	33	55	72	94	46	89	56	60	73	59	82	60
gemma-3-12b-it 12B	64.0%	64	33	55	83	100	50	84	44	50	60	64	64	50
nova-micro-v1	59.0%	59	33	47	67	100	50	79	33	70	73	36	64	50
mistral-nemo 12B	59.0%	59	17	53	67	94	46	79	44	40	67	45	64	70
ministral-3b-2512 3B	53.5%	54	50	45	67	67	38	84	33	50	60	32	64	70
lfm2-8b-a1b 8B/1B	0.0%	0	0	0	0	0	0	0	0	0	0	0	0	0

≥70% 40-70% <40%

Was wir bereits wissen

- Accuracy-Spannweite: **53–80 %**
- Stärke: Genetik
- Schwäche: Biomarker/ Metastasen

Möchten Sie Teil des BELLADONNA-Projekts werden?

Expert:innen für den Human Arm gesucht

- 40 Expert-Level MCQs
- Online in ca. 60 Minuten
- Auch in kurzen Sessions möglich
- Einladung und Startinfos per E-Mail

Jetzt Interesse registrieren:

Kontakt: catharina.wichmann@ukdd.de

Take-home message

1. **LLMs bereits Teil klinischer Kommunikation**
2. **Plausibel ≠ valide**
3. **BELLADONNA = Wissensbasis + Benchmarking + Validierung**

Vielen Dank für Ihre Aufmerksamkeit !

catharina.wichmann@ukdd.de

https://belladonna.kather.ai

Referenzen

1. Blease CR et al. Generative AI in primary care: online survey of UK GPs. BMJ Health Care Inform. 2024;31:e101102.
2. Boehm KM et al. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. Nat Commun. 2025;16:2106
3. Ferber D et al. GPT-4 for Information Retrieval and Comparison of Medical Oncology Guidelines. NEJM AI. 2024;1:AIcs2300235.
4. Lee Y, Ferber D, Kather JN et al. How AI agents will change cancer research and oncology. Nat Cancer. 2024;5:1765-1767
5. Ferber D et al. Autonomous AI Agents for Clinical Decision Making in Oncology. arXiv. 2024; arXiv:2404.04657.
6. Marra A et al. Artificial intelligence entering the pathology arena in oncology. Ann Oncol. 2025. doi:10.1016/j.annonc.2025.03.006.
7. GeneSilico Copilot. Improved precision oncology question-answering using agentic LLM. medRxiv. 2024. doi:10.1101/2024.09.20.24314076.
8. Hoeper C, Ferber D, Kather JN et al. ATheNa-Breast: Real-world pilot of an AI chatbot for breast cancer. JCO. 2025;43(16_suppl):e13623.
9. Singhal K et al. (Med-PaLM 2) Toward expert-level medical question answering with LLMs. Nature Medicine. 2023.
10. Lyu J et al. Hallucination incidence in LLM responses to oncology questions: meta-analysis. JCO. 2025;43(16_suppl):3866.
11. Bitterman DS et al. Promise and perils of large language models for cancer survivorship. JCO. 2024;42(14):1607-1611.
12. American Medical Association. Physician Survey on Augmented Intelligence. 2024
13. KFF. Poll: 1 in 3 Adults are Turning to AI Chatbots for Health Information, Equaling the share who uye social media for health. 2026