Toward Oncology Digital Twins: Leveraging LLMs for Automated Longitudinal Data Ingestion in Ovarian Cancer Care

Leen Houri[1⋆], Natyra Tahiri[1,4], Gaspar Voelker[2], Elina Arlanch[1,2], Marie Koechert[1,2], Jonatan Frank[1], Leonid Kuligin[1,5], Jannik Lübberstedt[3], Keno Bressem[3,6,7], Martin Boeker[1], Maximilian Tschochohei[1,5], and Jacqueline Lammert[1,2]

[1] Chair of Medical Informatics, Institute for AI and Informatics in Medicine, TUM University Hospital, TUM School of Medicine and Health, Technical University of Munich (TUM), Munich, Germany
[2] Department of Gynecology, TUM University Hospital, TUM School of Medicine and Health, Technical University of Munich (TUM), Munich, Germany
[3] Institute of Diagnostic and Interventional Radiology, TUM University Hospital, TUM School of Medicine and Health, Technical University of Munich (TUM), Munich, Germany
[4] Department of Internal Medicine III, TUM University Hospital, TUM School of Medicine and Health, Technical University of Munich (TUM), Munich, Germany
[5] Google Cloud, Munich, Germany
[6] National Center for Tumor Diseases West, Essen, Germany
[7] Institute of Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany
leen.houri@tum.de

Abstract. Multidisciplinary tumor board (MTB) protocols condense years of oncological history into dense narrative text, a rich but unstructured knowledge source. Converting them into machine-readable form is a prerequisite for oncology Digital Twins, yet remains labor-intensive. We present a self-hosted framework that maps ovarian carcinoma patient journeys onto a structured data model using an open-weight large language model (gpt-oss-120b) with schema-constrained generation. The pipeline runs entirely within the institutional network, so no patient data leaves the hospital. We extract patient journeys for 200 patients and evaluate fidelity on a 50-patient expert gold standard, using content-based field alignment robust to treatment-line ordering. Agreement is highest for well-defined structural fields (identifiers, staging, histology, document metadata), reaching 98–100%, and remains strong across categorical fields; sparse result-valued fields such as the biomarker value are the main remaining challenge. Clinician review raised overall field-level agreement to 92.5%. The gold standard is further validated against pharmacy records, which independently corroborate most of its dispensable drugs (recall 0.92, precision 0.84). We thereby establish a privacy-compliant

⋆ Corresponding author.

pathway toward structured substrates for medical Digital Twins, as an enabling data ingestion component rather than a Digital Twin itself.

Keywords: Digital Twin · Clinical Information Extraction · Large Language Models · Ovarian Cancer

1 Introduction

The vision of a Digital Twin in oncology, a computable model that simulates a patient's therapeutic trajectory and forecasts treatment response, is constrained by the quality of its underlying data [1, 6]. Such a model requires a chronological state-space representation capturing the entire evolutionary history of the tumor under prior therapies, yet this longitudinal information remains locked in the unstructured narrative text of Electronic Health Records (EHRs). Manual abstraction does not scale, and cancer registries rarely capture the cycle-by-cycle detail that individual simulation demands.

Multidisciplinary tumor board (MTB) protocols are an exceptionally dense source of this knowledge, condensing years of diagnostic and therapeutic history into a single narrative. The challenge is acute in ovarian cancer, where patients undergo surgical cytoreduction, sequential lines of systemic therapy, and biomarker-guided maintenance. While LLMs can extract discrete data points from clinical reports [2, 5], reconstructing an accurate multi-year timeline remains difficult: LLMs struggle with temporal reasoning in recurrence settings [4] and are prone to hallucinations [8]. We mitigate both by anchoring extraction to the single most recent protocol, which recapitulates the prior journey, and by constraining outputs to a schema with date invariants enforced in a validation loop. Compounding these technical limits, European data protection requirements mandate that ingestion pipelines run within institutional networks, ruling out proprietary cloud models in favor of locally hosted, open-weight alternatives.

To address this, we introduce a self-hosted framework with three parts: a structured data model for time-series oncological simulation; an open-weight LLM (gpt-oss-120b) with schema-constrained generation that maps narrative text onto this model; and evaluation against two references, an expert clinician gold standard and pharmacy dispensing records. By showing that locally hosted LLMs can parse complex disease trajectories with verifiable accuracy, we establish a privacy-compliant pathway toward reliable longitudinal structured data for medical Digital Twins. We therefore address the ingestion layer only: the pipeline produces the structured substrate required to build an oncology Digital Twin. Our contribution is threefold: a clinician-defined data model for longitudinal oncological state, a privacy-compliant extraction pipeline, and an evaluation methodology together with a consensus-built, externally corroborated reference standard for a setting in which reliable ground truth is scarce.

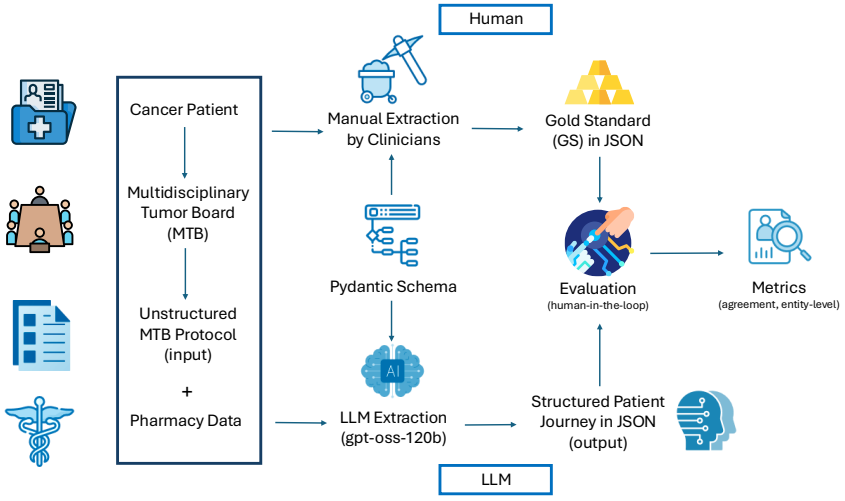

Fig. 1. Overview of the extraction and evaluation pipeline. The same MTB protocol is independently extracted into a structured patient journey by clinicians (gold standard) and by the model (gpt-oss-120b), both conforming to one Pydantic schema. Model output is compared field by field and entity by entity against the gold standard, with the automatic match/mismatch verdicts reviewed by a clinician; the gold standard is additionally validated against pharmacy dispensing records. Icons courtesy of Flaticon.com. Code at https://github.com/leenhouri/mtb-extraction-tool.

2 Methods

The pipeline reconstructs longitudinal patient journeys from gynecologic MTB protocols in three components (Fig. 1): (i) a Pydantic data model of the patient journey; (ii) a single-pass LLM extraction layer with schema-constrained generation and a bounded validation-retry loop; and (iii) a two-source evaluation framework that scores extractions against an expert gold standard and against institutional pharmacy records, using a content-based alignment so that clinically equivalent extractions are not penalized for differences in ordering or wording.

2.1 Patient Journey Data Model

The data model captures the patient journey at the granularity required to link therapeutic interventions to the clinical parameters that drive them. Each patient is represented hierarchically: a patient root (identifier, date of birth) links to an ordered list of documents, each carrying diagnoses, biomarkers, treatments, and tumor board outcomes (Table 1). Structuring the journey by document keeps every extracted fact attached to its source report and lets the same schema scale to multiple reports per patient, although the present study extracts a single protocol per patient. The model is a Pydantic schema supporting runtime validation

Table 1. The patient journey data model. Each patient links to an ordered list of documents carrying diagnoses, treatments, and tumor board outcomes. Categorical fields use controlled vocabularies defined with the clinical team; undocumented fields are recorded as *Not documented*.

Entity	Attributes	Controlled vocabulary
Patient	id, date of birth, documents	–
Document	type, date, diagnoses, treatments, tumor board outcomes	type: physician letter, tumor board protocol
Diagnosis	tumor entity, date, FIGO stage, TNM stage, resection status, recurrence, biomarkers	resection status: R0, R1, R2, complete, partial, no surgery, unclear; recurrence: yes, no, unclear
Biomarker	name, value, type, date	type: germline, somatic, unclear
Treatment	modality, line of therapy, status, start date, end date, medications, surgeries	modality: surgery, systemic, maintenance; status: completed/aborted, unclear
Medication	name, dosage, interval, start date, end date	–
Surgery	date, type, resection status	type: laparotomy, laparoscopy, other; resection status: R0, R1, R2, complete, partial, unclear

and machine-readable export, with controlled-vocabulary fields defined with our clinical collaborators.

2.2 Schema-Constrained Extraction

The extraction layer is built around gpt-oss-120b [7], an open-weight Mixture-of-Experts model served on institutional infrastructure through a vLLM server [3] with a LiteLLM proxy, so the full pipeline runs without transmitting patient data outside the institutional network. By institutional convention, the most recent gynecologic MTB protocol recapitulates the full prior treatment journey up to the meeting date, so a single document per patient suffices and avoids cross-document reconciliation.

Structured output is obtained through schema-constrained generation: the Pydantic schema is translated to a JSON schema and passed to the model via LangChain's structured-output interface, fixing field names, nesting, value types, and optionality. Controlled vocabularies are specified in the prompt rather than the schema, since the model adhered more reliably to enumerated values stated in the prompt body. The prompt also encodes domain knowledge, abbreviation expansions, German-to-English term mappings, therapy-line construction rules, and one validated worked example (one-shot). Generation and validation are orchestrated as a directed state graph (LangGraph): a validation node re-parses

each response with Pydantic and checks document-level invariants (e.g., a treatment end date not preceding its start), re-prompting with the structured errors up to three times before a hard failure. Inference ran at temperature 0 on a single high-memory GPU.

2.3 Evaluation

We evaluate against two independent references: an expert gold standard manually extracted by clinicians, and institutional pharmacy dispensing records. The model produced records for 200 patients; evaluation uses the 50-patient gold-standard subset. Both comparisons run through a shared scoring layer that normalizes values, aligns list-valued items by clinical content, and produces a candidate match/mismatch verdict per field. Because automated equality can diverge from clinical judgment, a clinician reviews and corrects these verdicts.

Gold standard construction. Five clinicians built the gold standard through a consensus process. Four independently extracted the 50 patient histories into the schema, blind to one another and to the model output; their extractions were consolidated into a single reference, which an independent rater reviewed against the source protocols and corrected, with residual disagreements resolved by discussion among all five. Because it was constructed without sight of the model output, the model-versus-reference comparison is not circular. This yields two references: the consolidated extraction before review (Human) and the reviewed gold standard (GS). Agreement with the reviewed reference is the primary evaluation; the difference between the two quantifies the review effect, tested per field with the exact McNemar test and Benjamini–Hochberg correction. Inter-rater agreement between two of the independent extractions is reported as percent agreement, Cohen's κ, and Gwet's AC1. Because full manual extraction and source verification by multiple clinicians is highly labor-intensive, we deliberately prioritized the quality of the reference standard over the size of the evaluated cohort.

Normalization and alignment. References and prediction pass through one symmetric normalization: dates to ISO format with partial dates resolved to a fixed convention, biomarker names canonicalized through an explicit synonym table (name equivalences only, never value-to-value mappings), and documented synonyms, units, and emptiness tokens folded onto canonical forms. Because treatments, medications, surgeries, and biomarkers are lists whose order is unreliable, list-valued items are aligned by content as a one-to-one bipartite matching (treatment lines by date proximity and shared active ingredients, medications by ingredient and start month, surgeries by date month, biomarkers by name), so genuinely distinct items are not collapsed.

Field-level agreement. An automated rule assigns each aligned field a candidate match or mismatch: two values match when they are equal under the

normalization above (including when both are absent), and count as a mismatch otherwise. A clinician then reviewed and corrected these candidate verdicts for both references over all 50 patients, because a strict equality rule can flag clinically equivalent values as mismatches; the reported agreement therefore reflects clinical meaning rather than surface string identity. Agreement is the proportion of matches, $a/(a+b)$, reported per schema field type and overall. We report the overall agreement with a 95% confidence interval from a cluster bootstrap that resamples the 50 patients (respecting within-patient dependence), and the prevalence-robust Gwet's AC1 alongside Cohen's κ for inter-rater agreement, since κ is deflated when one category dominates.

Entity-level precision and recall. Field-level agreement presupposes that an item was aligned at all, and therefore does not capture entities the model missed entirely or introduced spuriously. We therefore additionally report entity-level precision, recall, and F1 for the list-valued entities (treatments, medications, surgeries, biomarkers), derived from the same one-to-one content-based alignment: an aligned pair counts as a true positive, an unmatched predicted entity as a false positive, and an unmatched gold entity as a false negative.

Pharmacy concordance. As a model-independent check, we compare gold-standard medications against pharmacy dispensing records. Per patient and ingredient, dispensing events are aggregated into a span, active ingredients are canonicalized on both sides, and supportive medications not in the data model are excluded. We report presence (precision, recall, F1 against dispensed ingredients) and date concordance within ± 1 and ± 2 months; a tolerance is used because the date documented at the tumor board and the date a drug is actually dispensed can differ by weeks. Because the pharmacy dispenses intravenous drugs only, orally administered drugs (oral PARP inhibitors, oral endocrine therapy) are excluded. To prevent temporal leakage, each span is right-censored at the protocol date, since a gold standard derived from the protocol cannot reflect later dispensings.

Code Availability: Code available at https://github.com/leenhouri/mtb-extraction-tool. **Data Availability:** Available on request from the corresponding author; not public due to privacy restrictions. **Ethics:** Local ethics approval was granted (No. 2025-362-S-CB).

3 Results

The pipeline was evaluated on the 50-patient gold standard: per-field extraction agreement against the reviewed reference (Sec. 3.1), entity-level precision and recall (Sec. 3.2), and validation of that reference against pharmacy records (Sec. 3.3).

Table 2. Per-field agreement of the LLM extraction against the reviewed gold standard (GS) and the pre-review consolidation (Human). Δ = GS − Human; q is the Benjamini–Hochberg-adjusted exact-McNemar p-value ($^{*}q < .05$, $^{**}q < .01$, $^{***}q < .001$), indicative given within-patient dependence. The overall GS agreement is reported in the text with a 95% cluster-bootstrap interval.

Category	Field	n	GS	Human	Δ (pp)	q
Patient	ID	50	100.0	98.0	+2.0	1.000
	Date of birth	50	100.0	94.0	+6.0	0.536
Document	Type	50	100.0	100.0	+0.0	1.000
	Date	50	100.0	100.0	+0.0	1.000
Diagnosis	Tumor (histology)	51	100.0	98.0	+2.0	1.000
	Date	51	90.2	84.3	+5.9	0.625
	FIGO stage	51	98.0	98.0	+0.0	1.000
	TNM stage	51	100.0	100.0	+0.0	1.000
	Resection status	51	100.0	90.2	+9.8	0.156
	Relapse	51	96.1	88.2	+7.8	0.543
Biomarker	Name	97	90.7	84.5	+6.2	0.543
	Type	97	97.9	68.0	+29.9	$<0.001^{***}$
	Value	97	79.4	79.4	+0.0	1.000
	Date	97	87.6	71.1	+16.5	0.034^{*}
Treatment	Type	262	95.4	85.9	+9.5	0.002^{**}
	Start date	262	90.5	80.5	+9.9	0.007^{**}
	End date	262	90.1	81.3	+8.8	0.017^{*}
	Status	262	93.9	72.9	+21.0	$<0.001^{***}$
	Treatment line	262	91.6	78.6	+13.0	$<0.001^{***}$
Medication	Name	284	93.3	87.0	+6.3	0.061
	Dosage	284	96.5	98.2	-1.8	0.543
	Interval	284	94.4	99.6	-5.3	0.002^{**}
	Start date	284	86.3	81.7	+4.6	0.457
	End date	284	83.8	83.1	+0.7	1.000
Surgery	Type	111	91.9	62.2	+29.7	$<0.001^{***}$
	Date	111	95.5	82.0	+13.5	0.003^{**}
	Resection status	111	94.6	91.9	+2.7	0.917
Tumor board	Date	50	96.0	94.0	+2.0	1.000
	Input	50	100.0	98.0	+2.0	1.000
	Recommendation	50	96.0	98.0	-2.0	1.000
Overall		**4,107**	**92.5**	**85.3**	**+7.2**	$\mathbf{<0.001^{***}}$

3.1 Extraction Performance

Table 2 gives per-field agreement against the reviewed gold standard (GS) and the pre-review consolidation (Human), the difference (Δ), instances (n), and the Benjamini–Hochberg FDR-adjusted McNemar q-value; because fields within a patient are not independent, q is indicative rather than strict.

Against the reviewed gold standard the LLM extraction reaches 92.5% agreement over 4,107 fields (95% CI 90.9–94.2). The structural and categorical back-

Table 3. Entity-level precision, recall, and F1 over the 50-patient gold standard, derived from the one-to-one content-based alignment. TP: aligned pairs; FP: predicted entities without a gold counterpart; FN: gold entities without a predicted counterpart.

Entity	TP	FP	FN	Precision	Recall	F1
Treatment	256	0	6	1.00	0.98	0.99
Medication	265	2	17	0.99	0.94	0.97
Surgery	98	0	4	1.00	0.96	0.98
Biomarker	88	0	9	1.00	0.91	0.95
Overall	707	2	36	1.00	0.95	0.97

bone is recovered almost exactly: identifier, date of birth, document type and date, tumor histology, TNM stage, and tumor board input reach 100%, FIGO stage 98.0%. Agreement is lowest on sparse fields, in particular the biomarker value (79.4%) and the medication end and start dates (83.8% and 86.3%). Since jointly absent fields count as matches (20.1% of all fields), restricting to fields documented on at least one side lowers overall agreement to 90.7%.

Clinician review of the gold standard raised the measured agreement substantially: overall agreement against the reviewed reference exceeded that against the pre-review consolidation by 7.2 pp (85.3% to 92.5%, $q < 0.001$). The largest gains concentrate in the interpretive fields where the original consolidation was least reliable (biomarker type, surgery type, treatment status, treatment line; all $q < 0.05$, Table 2); only medication interval moved in the opposite direction (-5.3 pp, $q = 0.002$), where the review revised the reference away from a value the model and consolidation shared.

3.2 Entity-Level Performance

Table 3 reports entity-level precision, recall, and F1 for the list-valued entities, computed from the same content-based alignment used for field scoring. This complements field-level agreement by quantifying entities that were missed or spuriously introduced rather than only those that were misfilled.

Entity recovery is near-complete across all four entity types (overall F1 0.97 over 743 gold and 709 predicted entities). Spurious entities are almost absent: only two medications were extracted without a gold counterpart, and no treatment line, surgery, or biomarker was invented (overall precision 0.997). The residual errors are omissions rather than fabrications, concentrated in medications (17 missed) and biomarkers (9 missed), which are also the entities most often mentioned only in passing in the protocol text. Taken together with the field-level results, this locates the remaining error mass in the *values* of correctly identified entities, in particular the biomarker value and the fine-grained medication dates, rather than in the presence or absence of the entities themselves.

Table 4. Concordance of the gold standard with pharmacy dispensing records (50 patients), restricted to drugs the pharmacy dispenses (oral PARP inhibitors and endocrine therapy excluded) and to dispensings up to the protocol date. Presence compares extracted against dispensed active ingredients; date concordance is over matched ingredients.

Measure	Value
Presence precision	0.84
Presence recall	0.92
Presence F1	0.88
Start date concordance ($\leq$2 months)	0.75
Start date concordance ($\leq$1 month)	0.67
End date concordance ($\leq$2 months)	0.78
Start & end concordance ($\leq$2 months)	0.55

3.3 Validation of the Gold Standard

The gold standard was validated against pharmacy dispensing records for all 50 patients (Table 4). Presence recall was 0.92 and precision 0.84 (F1 0.88), the lower precision consistent with drugs dispensed outside the institutional pharmacy, canonicalization mismatches, or timing differences. Among matched ingredients, start dates were concordant within two months in 0.75 of cases and end dates in 0.78, indicating the reference captures the therapy trajectory well. Inter-rater agreement between two of the independent extractions was 86% (Cohen's $\kappa = 0.84$, Gwet's AC1 = 0.85); on a few high-prevalence categorical fields κ is deflated by the dominance of one category, so AC1 is the fairer summary there. This human-to-human estimate characterizes the difficulty of the extraction task itself; it is not measured on the same scale as the model-to-gold agreement and should not be read as evidence that the pipeline outperforms individual clinicians.

4 Discussion

Against the reviewed gold standard the LLM-based extraction reached 92.5% overall, with a clear gradient: the structural and categorical backbone (identifiers, document and diagnosis fields, staging, resection and relapse status, biomarker and tumor board fields) was recovered at or near complete agreement, while interpretive and temporal fields were lower and the biomarker value (79.4%) hardest. For a feasibility study this is the central result: the structural core of the patient history can be reconstructed reliably without sending data outside the institutional network, while a few sparse, result-valued fields define the open challenges. The entity-level analysis (Table 3) sharpens this picture: whole entities are recovered at F1 0.97 with only two spurious medications and no invented treatment line, surgery, or biomarker across the 50 patients. The clinically most consequential failure mode for a downstream twin, a fabricated or

silently dropped therapy, is therefore rare, and the residual disagreement lies in the values recorded for entities that were themselves correctly identified.

A second finding compares the model against the reference before and after review: it agreed with the reviewed gold standard at 92.5% but with the pre-review consolidation at only 85.3% ($\Delta = +7.2$ pp, $q < 0.001$). Because the gold standard was built without sight of the model output, this gap is informative: where the consolidation and corrected reference disagreed, the model more often matched the source-verified value than the single-pass consolidation did (notably biomarker type, surgery type, treatment status, and treatment line). Beyond its effect on the reported numbers, this observation carries a methodological implication: measured extraction performance in this domain depends materially on how the reference itself is constructed, and single-pass manual extraction is not a neutral reference standard. Several schema fields, in particular the tumor board input and the surgery outcome, were initially interpreted differently by trained clinicians, and only explicit field definitions, agreed across annotators, made the comparison meaningful.

A particular strength is the gold standard itself: independently extracted by four clinicians, finalized by consensus review against source, and corroborated by the pharmacy records (92% of dispensable drugs appear in the dispensing data), so the reference is externally validated rather than a lone expert opinion. The weakest field, the biomarker value, has an addressable cause: the prompt elicited a qualitative status where the gold standard recorded the exact quantitative result.

Limitations. Several limitations bound the scope of these results. The study covers a single institution, a single tumor entity, and a single document type, and the quantitative evaluation rests on a 50-patient reference subset of the 200-patient extraction cohort; the evidence is therefore not sufficient to support claims of general applicability across oncology. The single-document design relies on the institutional convention that the most recent protocol recapitulates the prior journey. The pharmacy table omits orally administered drugs, which makes presence precision a lower bound. Raw agreement counts two absent values as a match, which is why we also report agreement restricted to documented fields (90.7%, Sec. 3.1). Finally, the pipeline is evaluated with one on-premise open-weight model: because our clinical deployment currently serves a single model under institutional data-governance constraints, we do not compare against other model families, smaller model scales, or rule-based extraction baselines, and we do not report a component-level ablation isolating the contributions of schema-constrained generation, controlled vocabularies, prompt-encoded domain knowledge, and the validation-retry loop. The results should accordingly be read as a single validated reference point rather than a demonstrated state of the art.

5 Conclusion and Outlook

We presented a pipeline that reconstructs longitudinal ovarian carcinoma patient journeys from MTB protocols with an open-weight LLM hosted entirely within the institutional network, reaching 92.5% agreement against a reviewed expert gold standard while the biomarker value and fine-grained temporal fields remain the principal challenge. Future work includes scaling to the full cohort and to several reports per patient (already supported by the data model), and evaluating further tumor entities to test generality. A systematic comparison across model families and model scales, including the smallest model that still reaches clinically acceptable performance, together with a controlled ablation of the individual pipeline components and a comparison against rule-based extraction baselines, is the central aim of our follow-up study; establishing these would indicate how far the approach transfers to institutions with more constrained hardware. Together these steps would move the pipeline from a single-entity feasibility study toward a scalable substrate for oncology Digital Twins.

Acknowledgments. This study was funded by the Bavarian State Ministry of Health, Care and Prevention (Bayerisches Staatsministerium für Gesundheit, Pflege und Prävention; grant number DGP-2024-01), with Bayern Innovativ GmbH as project management agency.

Disclosure of Interests. J.L. received honoraria from the Forum for Continuing Medical Education (FomF), AstraZeneca GmbH (Germany), and Novartis Pharma GmbH (Germany). The remaining authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmad, S., Bartelt, B., Enzmann, M., Kohlhammer, J., Wesarg, S., Wolf, R.: Secure Medical Digital Twins: A Use-Case Driven Approach. In: International Workshop on Digital Twin for Healthcare 2025 (2025). https://doi.org/10.24406/publica-5836
2. Bartels, S., Carus, J.: From text to data: open-source large language models in extracting cancer related medical attributes from German pathology reports. Int. J. Med. Inf. **203**, 106022 (2025). https://doi.org/10.1016/j.ijmedinf.2025.106022
3. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles, pp. 611–626 (2023)
4. Lammert, J., Pfarr, N., Kuligin, L., Mathes, S., Dreyer, T., Modersohn, L., Metzger, P., Ferber, D., Kather, J.N., Truhn, D., Adams, L.C., Bressem, K.K., Lange, S., Schwamborn, K., Boeker, M., Kiechle, M., Schatz, U.A., Bronger, H., Tschochohei, M.: Large language models-enabled digital twins for precision medicine in rare gynecological tumors. npj Digit. Med. **8**, 420 (2025). https://doi.org/10.1038/s41746-025-01810-z
5. Langhout, S.A.M., Hermans, S.J.F., Smit, A.J.T., Berkx, E., Kurk, S.A., Schade, K.J., Posthuma, E.F.M., Visser, O., Cornelissen, J.J., Huijgens, P.C., Versluis, J.,

van der Wilt, M., Dinmohamed, A.G.: Real-time data in cancer registries: validation of an automated data extraction system. iScience **28**(8), 113056 (2025). https://doi.org/10.1016/j.isci.2025.113056
6. Olawade, D.B., Oisakede, E.O., Bello, O.J., Analikwu, C.C., Egbon, E., Ojo, A.: Digital twins in oncology: from predictive modelling to personalised treatment strategies. Crit. Rev. Oncol. Hematol. **220**, 105171 (2026). https://doi.org/10.1016/j.critrevonc.2026.105171
7. OpenAI: gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925 (2025)
8. Xu, Z., Jain, S., Kankanhalli, M.: Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817 (2025)