Evaluating Temporal Context Strategies: Longitudinal Tumor Extraction in Hepatocellular Carcinoma

A. Altar Lüser[1], Yuxuan Chen[1], Imge Yüzüncüoglu[1], Philippe Thomas[1], Roland Roller[1], Robert Oehring[2], Nikitha Shruthi Ramasetti[2], Felix Krenzien[2], Sebastian Möller[1,3],

[1]DFKI Berlin, Germany [2]Charité – Universitätsmedizin Berlin, Germany
[3]Technische Universität Berlin, Germany

Correspondence: ahmet_altar.lueser@dfki.de

Abstract

Tumor board decisions often depend on information that is spread across several clinical documents written at different time points. In this work, we study zero-shot extraction of currently active hepatocellular carcinoma tumor findings from longitudinal German clinical records. We evaluate five open-weight LLMs on 114 tumor board cases and compare inputs based on the latest reports, the available longitudinal report history, and retrieved parts of the patient history. We focus on tumor size and tumor location as they are clinically important for describing tumor burden and are often mentioned repeatedly, change over time, or become inactive after treatment. Our results show that providing the available longitudinal report history chronologically gives the strongest overall performance in this dataset, while retrieval with temporal focus improves over purely semantic retrieval but is limited by missing follow-up or diagnostic context in the retrieved chunks. The main bottleneck is deciding whether a previously described tumor should still be counted as active after later treatment, resection, local control, residual vitality, or follow-up evidence.

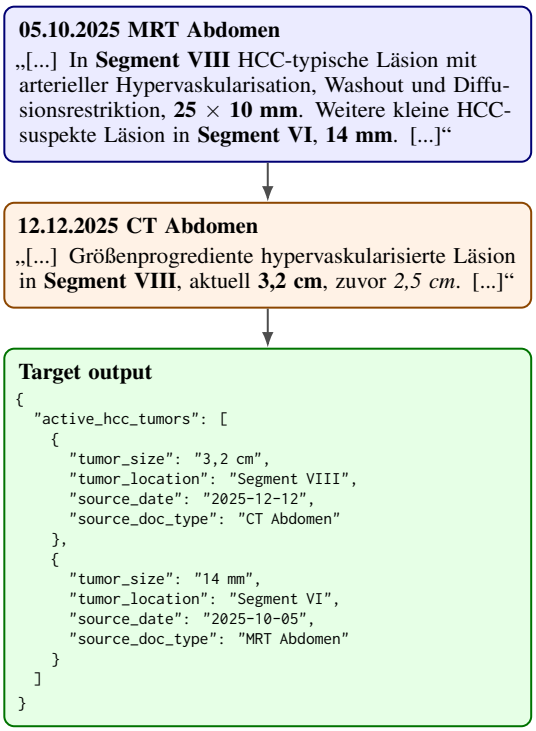

Figure 1: Example of active tumor extraction from longitudinal reports. The later CT updates Segment VIII, while Segment VI remains active without later contradictory evidence.

1 Introduction

Clinical decisions in oncology often rely on tumor board discussions, where a multidisciplinary team of physicians reviews a patient's history across multiple documents, including radiology reports, pathology findings, discharge summaries, and operative notes (Patkar et al., 2011; Okasako and Bernstein, 2022). These documents are written at different time points and describe the patient at different stages of diagnosis, treatment, and follow-up. In this work, we focus on predicting tumor findings that describe the patient's active disease at the time of the tumor board session.

Identifying currently active tumor findings from a patient history is challenging because patients may have multiple lesions at the same time, and the same lesion may be described repeatedly across reports **with changes in size, viability, treatment response or absence**. Measurements can also vary across imaging modalities (e.g., MRI, CT) and reporting practices, so a changed value does not necessarily indicate tumor progression or a separate lesion (Muenzel et al., 2012; Armbruster et al., 2021). CT and MRI can also disagree on whether a lesion shows HCC-typical features, so the same lesion may be described differently depending on the modality used (Chernyak et al., 2018). Conversely, an earlier lesion may remain relevant if no later evidence rules it out. The challenge is therefore to decide which tumor mentions still describe an active disease after considering the patient history (Yim et al., 2016; Datta et al., 2022). Figure 1 illus-

trates this problem with a simplified longitudinal example.

Explicitly modeling these decisions can become rule-heavy and task-specific, and may not generalize well across settings. Classical named entity recognition (NER) models can be used for parts of this problem, such as detecting tumor-size and tumor-location mentions, but the full task is not a standard local span-extraction problem. Addressing it with NER would require either additional steps for linking repeated mentions across reports and resolving which findings belong to the final case-level output, or task-specific annotations for learning these document-level decisions.

To address this challenge, we study whether large language models (LLMs) can offer a more flexible alternative for extracting active tumor findings from longitudinal clinical context using prompt-based instructions (Agrawal et al., 2022; Sivarajkumar et al., 2024). For LLM-based extraction, performance depends on which parts of the patient history are provided to the model. Using only the most recent reports may omit earlier findings that remain clinically relevant, whereas providing the full chronological history may introduce outdated measurements, treated lesions, or repeated descriptions that no longer reflect the current disease state. Since LLMs do not always use long contexts robustly (Liu et al., 2024), we evaluate temporal context strategies through five input variants: the most recent report of each report type; two chronological-history variants, using either the full patient history or reports grouped by report type; and two retrieval-based variants, selecting chunks either by semantic similarity alone or with temporal reranking, which incorporates time distance and samples evidence from earlier, middle, and later parts of the patient history.

We analyze zero-shot extraction of currently active tumor findings from longitudinal German clinical documents in hepatocellular carcinoma (HCC) (Vogel et al., 2022). We focus on tumor size and location as they describe the extent and distribution of tumor burden, information relevant for staging, treatment recommendation, and follow-up assessment in HCC (Reig et al., 2022). We use open-weight language models in the 8B to 70B parameter range, reflecting practical constraints in clinical settings where privacy considerations may limit the use of proprietary systems (Zhong et al., 2025).

Our **contributions** are: (1) we evaluate how much patient history is needed for extracting active tumor findings, comparing latest-report, full-history, and retrieval-based inputs; (2) we test whether retrieval improves when semantic relevance is combined with temporal reranking and temporal selection constraints; and (3) we analyze why retrieval-based inputs remain less precise than full-history inputs. We find that providing the available report history gives the strongest overall performance in this dataset. Time-aware retrieval improves over semantic retrieval alone, but retrieved chunks can lack the follow-up context needed to decide whether older tumor findings remain active. Most errors involve failures to apply the criteria to decide whether a tumor is still active or not.

2 Related Work

Clinical Information Extraction Clinical information extraction (IE) has been studied with rule-based systems, classical machine learning (ML), transformer-based encoders, and more recently generative LLMs. Rule-based systems often use regular expressions, dictionaries, and manually written rules to extract structured information from free text (Friedman et al., 2004; Savova et al., 2010). These methods can work well when the target entity is clearly defined and expressed consistently, but they are rule-heavy and difficult to adapt when terminology, report style, language, or clinical parameters change. Transformer-based encoders can perform well when fine-tuned on in-domain annotated data, but their performance can drop under cross-corpus or unseen-domain evaluation; Kühnel and Fluck (2022) report an average 19% F1 drop in disease NER cross-evaluation. This is especially relevant for German clinical IE, where public annotated datasets are scarce and data protection requirements make dataset creation difficult (Hofenbitzer et al., 2025).

Generative LLMs are increasingly evaluated for clinical IE as extraction criteria can be specified in natural language without task-specific pretraining or fine-tuning. Agrawal et al. (2022) and Sivarajkumar et al. (2024) show that LLMs can perform zero- and few-shot clinical IE. For German medical IE, Spiegel et al. (2025) evaluate 7B LLMs on public German clinical datasets and find strong drug extraction performance but weaker diagnosis and treatment extraction. Clinical IE is also relevant for tumor board workflows, where structured information about tumor burden, location, treatment history, and current disease activity is needed for

Split	Patients	Cases	Docs	Avg docs/case	Avg token/case	Median tok.	P75 tok.	Median days diff	>8k	>16k
Dev	17	45	417	9.27	8084.51	6656.0	11340.0	310	44.44	8.89
Test	28	69	751	10.88	8612.25	8368.0	12192.0	266	50.72	11.59
All	45	114	1168	10.25	8403.93	7224.5	12166.75	285	48.25	10.53

Table 1: Dataset statistics by split. Median tok. and P75 tok. are the median and 75th percentile of the per-case token count. Median days difference is the median span between the earliest and latest document per case; the last two columns show the percentage of cases exceeding long-context budgets.

case presentation and decision support (Macchia et al., 2022; Yüzüncüoglu et al., 2026).

Longitudinal Clinical Context with LLMs Some oncology parameters cannot be extracted reliably from isolated sentences or single reports. Tumor findings may be mentioned repeatedly, measured in different reports, treated, resected, or contradicted by later evidence. Earlier work addressed this problem through tumor reference resolution and cross-document coreference (Yim et al., 2016; Datta et al., 2022). Recent work investigates whether LLMs can reduce the need for task-specific clinical IE pipelines. Builtjes and Hering (2026) show promising results for locally deployable open-source LLMs in longitudinal tumor extraction, achieving over 93% attribute-level accuracy on paired Dutch radiology reports. However, even long-context LLMs may fail to use relevant information reliably depending on its position in the input (Liu et al., 2024).

Retrieval-augmented generation (RAG) provides selected evidence instead of full reports, usually through vector-based retrieval (Lewis et al., 2020) or structured representations such as knowledge graphs (Wu et al., 2025a), with the goal of reducing noise and grounding LLM outputs. Xiong et al. (2024) report that MedRAG improves the accuracy of six LLMs by up to 18% over chain-of-thought prompting. Li et al. (2026) transform longitudinal electronic health records into patient-specific temporal knowledge graphs and align patient trajectories with clinical guideline graphs for cancer decision support. For active tumor extraction, retrieval-based inputs need to capture both tumor descriptions, such as size and location, and later status evidence, such as treatment response, progression, resection, or absence of residual disease. Otherwise, models may incorrectly treat outdated or treated findings as currently active. More generally, RAG remains sensitive to retrieval quality. When the retrieved context is irrelevant, incomplete, or incorrect, LLMs may produce misleading outputs, and generated medical claims may still be unsupported by retrieved or cited sources (Sohn et al., 2025; Wu et al., 2025b). Recent work on LLM-based medical agents suggests one possible direction for addressing these limitations by combining retrieval with external tools, memory, multi-agent workflows, and human-in-the-loop oversight (Qiu et al., 2024; Hartsock et al., 2026).

3 Data

3.1 Dataset and Splits

We evaluate on a longitudinal clinical dataset of patients with HCC, a primary form of liver cancer. The dataset consists of pseudonymized German clinical records collected in routine care. Patients have one or more clinical cases, where each case corresponds to one tumor board episode and consists of multiple free-text medical reports. We treat each case as a separate instance, although cases from the same patient may be clinically related. The study was approved by the relevant ethics committee (EA4/169/22) for scientific use.

Report type	Count	%
Radiology	412	35.27
Discharge letters	337	28.85
Pathology	145	12.41
Surgery notes	77	6.59
Endoscopy	71	6.08
Others	72	6.16
LiMAx	54	4.62

Table 2: Overall report-type distribution. LiMAx refers to liver function capacity tests, which are used to assess functional liver reserve. Rare reports such as consult notes are merged into *Others*.

Overall, the dataset contains 1,168 free-text reports from 45 patients and 114 cases. Reports include radiology reports, discharge letters, pathology findings, surgical notes, endoscopy reports, and other oncology-related documentation, so tumor information is distributed across heterogeneous document types and time points. Cases include more than 10 reports on average and contain extended periods, with a median of approximately 285 days between the first and last docu-

Setup	History Scope	Retrieval	Re-ranking	Temporal Bias	Context Budget(s)
No-History	Latest per type	–	–	–	8k, 16k
History Concat	Full history	–	–	–	8k, 16k, 32k
Type-grouped Concat	Full history	–	–	–	8k, 16k, 32k
Semantic RAG	Partial	Top-k	Yes	None (relevance only)	8k
Time-Aware RAG	Partial	Top-k	Yes	Recency-weighted	8k

Table 3: Comparison of input construction strategies. Retrieval-based methods differ in whether temporal information is incorporated during ranking.

ment. Nearly half of all cases exceed 8k tokens, and more than 10% exceed 16k tokens, highlighting the challenge of constructing model input under limited context budgets. Detailed dataset statistics are shown in Table 1, and the report-type distribution is shown in Table 2.

We split the data by temporal availability. The earlier subset is used for development and later data for held-out evaluation. To prevent patient-level leakage, all cases from the same patient are assigned to the same split. The development set contains 45 cases from 17 patients, and the test set contains 69 cases from 28 patients.

3.2 Annotation

The annotations describe clinically relevant tumor findings that are considered active at the time of the tumor board session. Here, active means that an HCC tumor or HCC-suspicious lesion is still present or clinically relevant according to the available longitudinal evidence. Lesions that were resected, successfully treated without residual viability, or explicitly described as no longer present were not annotated as active findings. The annotations cover two tumor parameters: *tumor size* (TSIZE) and *tumor location* (TLOC). TSIZE refers to the reported measurement of an active tumor, including one-, two-, or three-dimensional measurements where available. TLOC refers to the anatomical location of the tumor as the liver segment.

Metric	Dev	Test	All
Active HCC %	73.33	66.67	69.30
Active cases	33	46	79
Avg TSIZE/active	1.55	1.57	1.56
Avg TLOC/active	1.48	2.09	1.84

Table 4: Active HCC prevalence and average lesion parameter density among active cases.

Annotation was performed at case level. Annotated TSIZE and TLOC values indicate which mentions belong to the final active tumor findings for each case, but the dataset does not provide explicit lesion-level links between sizes and locations. It also does not annotate all historical tumor mentions or assign temporal-status labels to individual mentions. The reference labels were based on tumor board case documentation and iteratively refined with a domain expert physician, who retrospectively reviewed the annotations for consistency across development and test sets. The dataset does not include formal independent double annotation; remaining ambiguity mainly concerns small suspicious lesions and post-treatment findings with unclear residual activity. Overall, 79 of 114 cases contain at least one active HCC finding. Active-case prevalence and average parameter counts are summarized in Table 4.

4 Methods

4.1 Task Formulation

Given an input text for a case, the model must predict the currently active HCC tumor findings at the time of the tumor board session. The input may consist of the latest reports, the concatenated document history, or retrieved text chunks, depending on the context construction strategy described in Section 4.2. The model then outputs a structured list of active tumor entries. Each entry contains a TSIZE, TLOC, source date, and source document type. If no active HCC tumor finding is present, the model should return an empty list. The overall task is to return only findings that remain active according to the available temporal evidence.

From the predicted tumor list, we derive three evaluation outcomes: **(1) case-level active tumor status (binary classification)**, where a case is tumor-positive if at least one active tumor entry is predicted; **(2) TSIZE**, evaluated by comparing the predicted size values with the gold size values per case, regardless of order; and **(3) TLOC**, evaluated in the same way and independently from TSIZE.

4.2 Context Construction

To study how context construction affects performance, we compare five strategies that differ in how patient history is selected and organized (Table 3). These fall into two groups: **direct-history methods** that provide raw document timelines and **retrieval-based methods** that select relevant parts of the patient history.

Direct-history methods vary in how much of the timeline is included and how documents are structured within the context budget. *No-History* includes only the most recent report per document type. *History Concat* concatenates the available reports in chronological order within the context budget, while *Type-grouped Concat* organizes documents by report type before concatenation. We also refer to these two concatenation strategies as *full-history input*, since both provide the available patient history. When the full history exceeds the context budget, older documents are truncated.

Retrieval-based methods construct the input by selecting relevant text chunks. We use *Semantic RAG* to denote the non-temporal retrieval baseline, where chunks are selected by content relevance without temporal weighting. The retrieval pipeline itself combines sparse, dense, and cross-encoder relevance signals. *Time-aware RAG* additionally incorporates temporal recency into the ranking and applies temporal selection constraints. Both retrieval-based methods share the same retrieval pipeline and differ only in the final scoring function and the application of temporal selection constraints.

4.3 Retrieval and Selection Pipeline

Retrieval-based methods share a common pipeline (Figure 2). Each document history is split into overlapping text chunks (1000 characters, two-sentence overlap) to preserve complete clinical statements (stage 2). Candidate chunks are retrieved using a hybrid approach combining sparse retrieval (BM25) and dense semantic retrieval (stage 3).[1] Results are merged via Reciprocal Rank Fusion and reranked with a cross-encoder to produce final relevance scores (stage 4).

The two retrieval strategies differ only in the final selection step (stage 5). In **Semantic RAG**, chunks are selected based on semantic relevance, with the final context consisting of ten tumor-related and three resection-related chunks to ensure coverage of both tumor evidence and treatment history. These values, along with chunk size, were selected based on development experiments. In **Time-aware RAG**, chunk selection combines relevance with temporal recency using:

$$\text{score}(c) = (1 - \alpha) \cdot \text{rel}(c) + \alpha e^{-\lambda \Delta t(c)} \quad (1)$$

where $\text{rel}(c)$ is the reranker score and $\Delta t(c)$ is the temporal distance to the reference time. We set $\alpha = 0.8$ and $\lambda = 0.001$ based on development experiments (Appendix B). To prevent over-reliance on recent documents, selection is constrained to preserve temporal diversity, including evidence from older, intermediate, and recent time points, while limiting the number of chunks per report. All hyperparameters can be found in Table 12.

[1] We ablate the sparse and dense components of this stage in Appendix A.4.

Model	Size	Reasoning	Medical
LLaMA-3.1-8B-Instruct	8B	–	–
Qwen2.5-32B-Instruct	32B	–	–
Qwen3-32B	32B	✓	–
MedGemma-27B-text-it	27B	–	✓
LLaMA-3.1-70B-Instruct	70B	–	–

Table 5: Overview of evaluated open-weight models. The model set covers different sizes, a reasoning-capable model, and one clinically fine-tuned model.

4.4 Prompting and Output Schema

The prompt instructs the model to extract only currently active HCC or HCC-suspicious lesions after considering the longitudinal patient history. Document dates are used to resolve temporal conflicts, with newer relevant evidence overriding older contradictory findings. The prompt also encodes document-type priorities, treating radiology reports and discharge letters as primary evidence for current tumor activity, while pathology reports mainly provide supporting information. Outputs follow a fixed JSON schema with one entry per active tumor, containing tumor size, tumor location, source date, and source document type. If no active tumor is present, the model returns an empty list. The full prompt template and expected JSON format are provided in Appendix C.

4.5 Models

We evaluate five open-weight LLMs covering different model sizes, reasoning behavior, and medical specialization (Table 5). The goal is not to provide a full model comparison, but to test

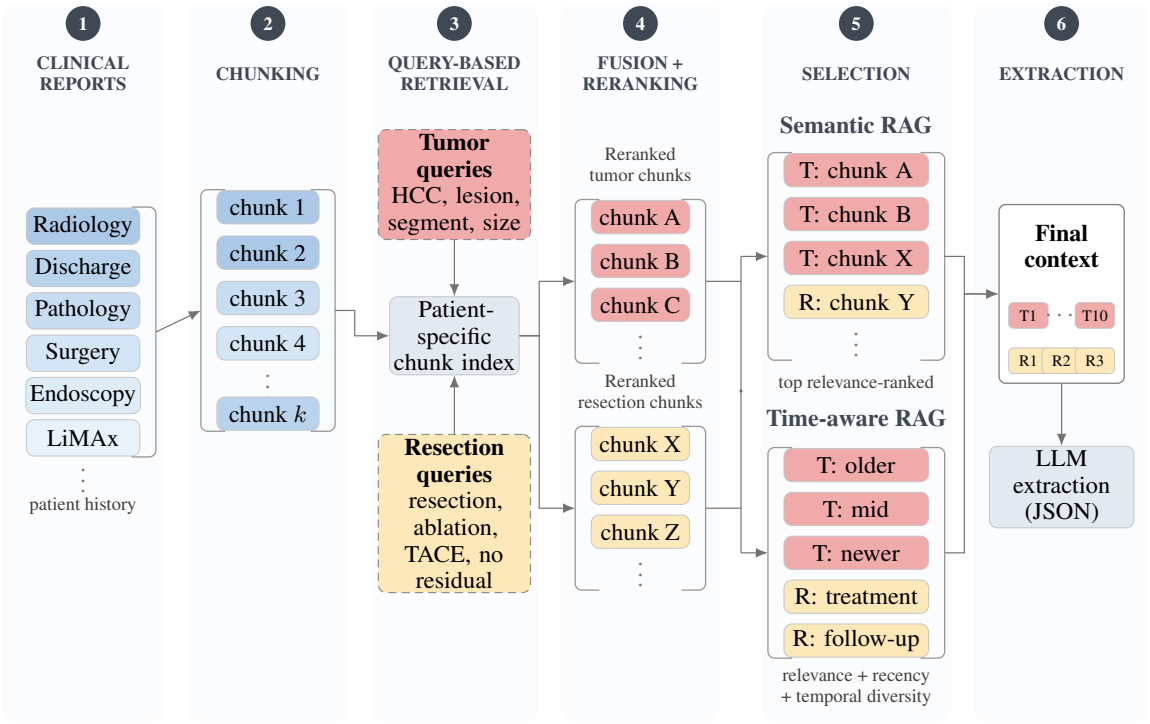

Figure 2: Retrieval and selection pipeline for the retrieval-based context construction strategies. Semantic RAG and Time-aware RAG share the same retrieval stages (1–4) and differ only in the final selection step (stage 5) before the extraction. Shading indicates the source document type in stages 1–2, and tumor (T) or resection (R) chunks thereafter. Query terms are abbreviated; the full templates are listed in Table 12.

whether the observed effects of context construction are consistent across different locally deployable LLMs. Our focus on open-weight models is motivated by the use of sensitive clinical patient data, since local deployment better aligns with privacy and data-governance constraints than proprietary API-based systems. The model set includes general instruction-tuned models from 8B to 70B parameters, a reasoning-oriented Qwen model, and MedGemma as a clinically fine-tuned model. Although MedGemma is primarily fine-tuned on English clinical data, it is built on Gemma 3 (Gemma Team et al., 2025) whose pretraining relies on multilingual data, and MedGemma is reported to maintain the general capabilities of its Gemma 3 base models (Sellergren et al., 2026), making it the closest domain-specialized option available for our German clinical setting at this scale.[2] All models are evaluated in a zero-shot setting with the same prompt, output schema, and decoding configuration.

[2]To separate the effect of medical fine-tuning from the effect of the underlying model, we additionally compare MedGemma-27B with its general-purpose counterpart Gemma-3-27B-it in Appendix A.3.

4.6 Evaluation Metrics

Performance is evaluated at case and parameter level. Case-level active tumor status is treated as binary classification, where a case is positive if at least one active HCC finding is predicted. At the parameter level, we report precision, recall, and F1 for TSIZE and TLOC. Since the annotations provide case-level TSIZE and TLOC values without lesion-level links, both parameters are evaluated independently as unordered sets per case. Before scoring, predictions are normalized for surface-form variation. For TSIZE, centimeter values are converted to millimeters, and multi-dimensional measurements are normalized dimension-wise, so equivalent forms such as *2.5 cm* and *25 mm* are matched. For multi-dimensional measurements, partial matches are accepted only if the predicted value matches the largest gold diameter after normalization. Approximate or rounded values, including those marked with *ca.*, are counted as incorrect unless they match the normalized gold value exactly. For TLOC, liver-segment labels are normalized across surface forms such as *Segment 8*, *Seg. 8*, and *S8*. Multi-segment mentions are split into separate segment labels, while broader anatomical descriptions such as *right liver lobe* are not scored.

Model	Strategy	Context	TSIZE				TLOC				Case			
			P	R	F1	σ	P	R	F1	σ	P	R	F1	σ
LLaMA-3.1-8B	no-history	8192	**0.560**	0.653	0.603	0.038	0.550	0.680	0.608	0.010	0.731	0.809	0.768	0.026
	type-grouped concat	8192	0.513	**0.819**	**0.631**	0.057	**0.597**	**0.825**	**0.693**	0.053	0.737	0.894	0.808	0.011
	history concat	8192	0.514	0.778	0.619	0.062	0.588	**0.825**	0.687	0.033	**0.741**	0.915	0.819	0.020
	semantic RAG	8192	0.350	0.694	0.465	–	0.490	0.742	0.590	–	0.727	0.851	0.784	–
	time-aware RAG	8192	0.340	0.722	0.462	–	0.500	0.794	0.614	–	0.733	**0.936**	**0.822**	–
MedGemma-27B	no-history	8192	0.573	0.764	0.655	0.021	0.632	0.742	0.682	0.030	**0.800**	0.851	**0.825**	0.020
	type-grouped concat	8192	**0.593**	**0.889**	**0.711**	0.049	**0.667**	**0.845**	**0.745**	0.028	**0.800**	0.851	**0.825**	0.003
	history concat	8192	**0.593**	**0.889**	**0.711**	0.025	0.661	**0.845**	0.742	0.018	**0.800**	0.851	**0.825**	0.007
	semantic RAG	8192	0.438	0.875	0.583	–	0.529	**0.845**	0.651	–	0.721	**0.936**	0.815	–
	time-aware RAG	8192	0.471	**0.889**	0.615	–	0.534	0.814	0.645	–	0.717	0.915	0.804	–
Qwen2.5-32B	no-history	16384	0.684	0.750	0.715	0.006	0.700	0.722	0.711	0.011	**0.837**	0.872	0.854	0.009
	type-grouped concat	8192	**0.718**	0.847	**0.777**	0.059	**0.748**	0.825	**0.784**	0.051	0.827	0.915	0.869	0.016
	history concat	8192	**0.718**	0.847	**0.777**	0.068	0.745	0.814	0.778	0.027	0.827	0.915	0.869	0.011
	semantic RAG	8192	0.504	0.833	0.628	–	0.588	0.825	0.687	–	0.742	**0.979**	0.844	–
	time-aware RAG	8192	0.543	**0.875**	0.670	–	0.632	**0.866**	0.730	–	0.807	**0.979**	**0.885**	–
Qwen3-32B	no-history	16384	0.743	0.722	0.732	0.033	0.683	0.711	0.697	0.029	0.844	0.809	0.826	0.035
	type-grouped concat	16384	**0.747**	0.861	**0.800**	0.030	**0.718**	0.814	**0.763**	0.013	**0.857**	0.894	**0.875**	0.034
	history concat	32768	0.678	0.819	0.742	0.004	**0.718**	0.814	**0.763**	0.021	0.837	0.872	0.854	0.030
	semantic RAG	8192	0.560	0.778	0.651	–	0.702	**0.825**	0.758	–	0.827	**0.915**	0.869	–
	time-aware RAG	8192	0.656	**0.875**	0.750	–	0.693	0.814	0.749	–	0.827	**0.915**	0.869	–
LLaMA3.1-70B	no-history	8192	0.591	0.764	0.667	0.037	0.617	0.763	0.682	0.019	0.827	0.915	0.869	0.006
	type-grouped concat	8192	0.630	**0.875**	0.733	0.067	0.702	**0.876**	0.780	0.021	0.818	**0.957**	0.882	0.013
	history concat	8192	**0.643**	**0.875**	**0.741**	0.042	**0.708**	**0.876**	**0.783**	0.014	**0.833**	**0.957**	**0.891**	0.013
	semantic RAG	8192	0.430	0.556	0.485	–	0.500	0.495	0.497	–	0.744	0.617	0.674	–
	time-aware RAG	8192	0.488	0.569	0.526	–	0.527	0.505	0.516	–	0.784	0.617	0.690	–

Table 6: Test-set performance of context construction strategies across models. Models are ordered by model size and strategies are shown in a fixed order. Bold marks the best precision/recall/F1 value within each model family for each metric; ties are bolded jointly. σ denotes variability across evaluated context lengths.

5 Results and Discussion

Table 6 shows the test-set results (development-set results are in Appendix D). Full-history input improves over the no-history setting mainly by recovering recall lost when only the latest reports are used. For TSIZE and TLOC, retrieval-based input usually does not outperform full-history input in F1, although time-aware RAG improves over semantic RAG. At the case level, retrieval is more competitive: for example, Qwen2.5-32B with time-aware RAG reaches 0.885 F1, close to the best case-level result of 0.891 from LLaMA-3.1-70B with history concatenation. For Qwen and MedGemma models, retrieval often preserves or slightly improves recall, but usually at the cost of lower precision.

Several configurations produce identical scores, which is expected as the difference between the two full-history variants is the order of the selected reports (chronologically or grouped by report type). Identical outputs therefore might indicate that a model is insensitive to document order, and differences appear only where a model is not. The clearest case is Qwen3-32B, where type grouping raises TSIZE F1 from 0.742 to 0.800.

Case-level detection is stable for a different reason. Because a case counts as positive when any active tumor is predicted, the decision depends on whether some tumor evidence survives context construction, not on which evidence. This also explains why the smallest model stays competitive at the case level while falling behind at the parameter level. LLaMA-3.1-8B is 6.9 F1 points below the best case-level result but 16.9 points below the best TSIZE result.

Since the main question is how context construction affects extraction, we do not analyze model-family differences in detail. Still, parameter count alone does not explain performance. Moving from 8B to 32B always improves results, but the newer 32B Qwen models are stronger than LLaMA-3.1-70B in several TSIZE and TLOC settings. The LLaMA models also gain less from retrieval-based input, suggesting that retrieval performance depends not only on the selected evidence but also on how well the model uses compact retrieved context.

Additional analyses of chronological versus type-grouped concatenation and context-budget effects are provided in Appendix A.1 and A.2. Type grouping gives small and inconsistent gains over chronological concatenation, and increasing the context budget gives limited benefit in this dataset.

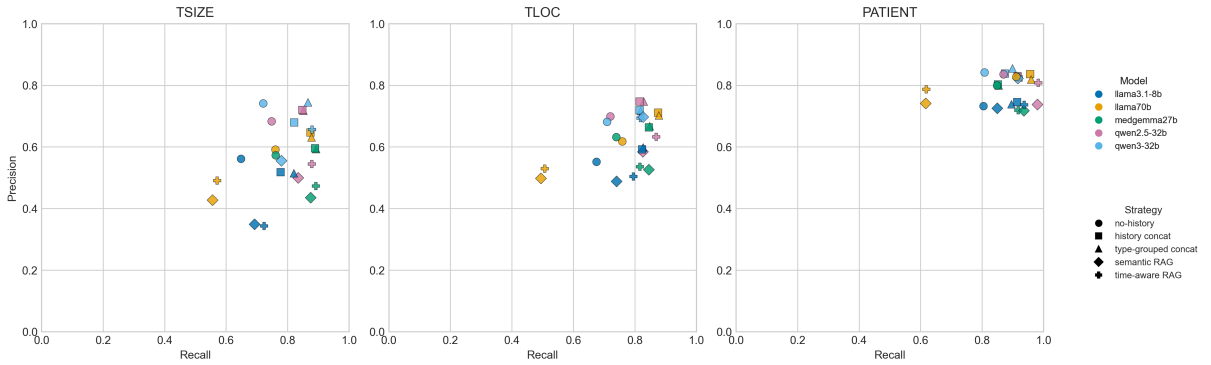

Figure 3: Test-set precision–recall trade-offs across strategies for TSIZE, TLOC, and case-level tumor detection. A zoomed-in view is shown in Figure 6.

5.1 Effect of longitudinal history

Adding longitudinal history mainly improves recall. Compared with no-history input, the best full-history variant improves recall for every model, by 9.7–16.6 points for TSIZE and 10.3–14.5 points for TLOC. This shows that latest reports alone often miss tumor parameters that remain relevant at the tumor board time point: a size or location may be mentioned in an earlier report and remain active unless later evidence rules it out, while later reports are needed to identify treated, resected, or no longer visible lesions. The recall gain usually comes with little precision loss. Compared with no-history input, the best full-history variant changes TSIZE precision by -4.7 to +5.2 points and TLOC precision by +3.5 to +9.1 points, with TSIZE for the smallest model as the main exception. The gains are larger for parameter extraction than for case-level tumor detection: full-history input improves case-level F1 by at most 5.1 points, while TSIZE F1 improves by 2.8–7.4 points and TLOC F1 by 6.3–10.1 points. Case-level detection is therefore easier, because a model can correctly mark a case as tumor-positive while still missing a lesion, selecting an outdated size, or returning an incomplete location. Overall, the results suggest that most models can use the additional history without being overwhelmed by historical mentions. The two full-history variants behave similarly, so we treat them jointly here and report the detailed comparison in Appendix A.1.

5.2 Retrieval-based input and temporal selection

Time-aware RAG improves over semantic RAG, especially for TSIZE. TSIZE recall increases for every model, with gains between 1.3 and 9.7 points, and TSIZE F1 improves in all models except MedGemma-27B. The largest gain appears for Qwen3-32B, where TSIZE F1 increases from 0.651 to 0.750. TLOC shows a weaker pattern. Time-aware RAG improves TLOC F1 for three of five models, with the largest gain for Qwen2.5-32B, from 0.687 to 0.730.

However, time-aware RAG remains weaker than full-history input in F1 because it is less precise. Compared with the best full-history variant, time-aware RAG reduces TSIZE precision by 9.1–17.5 points and TLOC precision by 2.5–18.1 points. For Qwen and MedGemma models, TSIZE recall is preserved or slightly improved under time-aware RAG, but the precision loss usually lowers F1. The LLaMA models benefit less from retrieval, with lower recall as well as lower precision.

This reflects a limitation of chunk-level retrieval for this task. Retrieved chunks may contain plausible tumor evidence, such as an older size or location, while omitting later reports that determine whether the lesion was treated, resected, locally controlled, or no longer viable. Retrieval success therefore cannot be defined only as finding a tumor measurement; the selected evidence also needs to include follow-up information that resolves current activity. Figure 3 shows the same pattern: time-aware RAG often shifts results toward higher recall, but retrieval-based methods remain less balanced than full-history input. This limitation is not caused by the choice of retriever. Ablating the sparse and dense components of the retrieval stage changes TSIZE F1 by at most 0.7 points and does not close the gap to full-history input (Appendix A.4).

The strongest results still leave room for improvement, with best F1 scores of 0.891 for case-level detection, 0.8 for TSIZE, and 0.784 for TLOC. We therefore analyze the remaining error patterns

in Section 6.

6 Error Analysis

We manually analyzed TSIZE errors for the Qwen models under history concatenation, semantic RAG, and time-aware RAG. The dominant error source was deciding whether a previously described tumor should still be counted as active after later treatment, resection, or follow-up evidence. Averaged across the three analyzed strategies, post-treatment interpretation accounted for 52% of TSIZE errors for Qwen2.5-32B and 29.2% for Qwen3-32B, suggesting that the reasoning model reduces, but does not eliminate, failures in applying temporal and treatment-aware decision rules.

Retrieval-based settings also showed context-gap errors. In some cases, retrieval selected chunks with older tumor information but failed to retrieve later evidence on treatment response, resection, local control, or residual vitality. In other cases, the retrieved context was too narrow and omitted surrounding diagnostic information needed to determine whether a finding corresponded to an HCC tumor. These errors help explain why time-aware RAG improves over semantic RAG, but still remains less precise than full-history input. The lower share of missing status evidence errors for Qwen3-32B suggests that stronger reasoning can sometimes compensate for incomplete or noisy retrieved context, although it does not remove the underlying retrieval limitation. Other errors involved outdated measurements, suspicious satellite lesions, rounded or hallucinated values, and malformed outputs. A more detailed error analysis is provided in Appendix A.5.

7 Conclusion

We evaluated temporal context strategies for zero-shot extraction of active HCC tumor findings from longitudinal German clinical records, where active findings refer to HCC or HCC-suspicious lesions that remain clinically relevant at the tumor board time point rather than all historical tumor mentions. **(1)** Providing the available report history chronologically gives the best performance in this dataset, likely because most patient histories fit within moderate context budgets and because the complete timeline includes status-resolving evidence that retrieval may fail to retrieve. Compared with using only the latest reports, this setting recovers more tumor size and tumor location evidence, showing that active tumor parameters are often distributed across the patient history. **(2)** Temporal retrieval improves over purely semantic retrieval, especially for tumor size recall, but does not outperform chronological full-history input in F1 because it remains less precise. Retrieved chunks may contain older tumor evidence without the later follow-up information needed to decide whether the finding is still active, and may also omit surrounding diagnostic context needed to determine whether a finding corresponds to an HCC tumor. **(3)** Main errors show that interpreting active tumor status remains a key bottleneck. Models need to better link tumor findings to treatment history, resection, local control, residual vitality, and post-therapeutic changes. Overall, open-weight models still need improvement for longitudinal active-tumor extraction.

8 Limitations

This study has several limitations. The annotations describe the final active HCC findings at the tumor board time point, but they do not track how each lesion changed over time. They also do not label each mention as current, past, treated, or resolved. As a result, older or treated findings are handled only through the final interpretation of active disease. The annotations were not independently double-annotated, and we therefore cannot report inter-annotator agreement. However, the reference labels are based on real tumor board case documentation, where the main clinically relevant HCC findings were part of the routine clinical review. The main source of annotation uncertainty is therefore not the primary tumor board finding, but smaller HCC-suspicious lesions, historical mentions without clear follow-up status, and post-treatment lesions where residual activity is ambiguous. To reduce this uncertainty, all annotations were retrospectively reviewed and refined with a domain expert physician. Future work should nevertheless include independent double annotation or adjudication to quantify remaining uncertainty.

TSIZE and TLOC are also not linked to the same individual lesion. We therefore evaluate them as unordered sets rather than checking whether each size is matched to the correct location. This is sufficient for the current tumor board setup, but it does not test this more detailed association problem. We also do not evaluate the source date and source document type requested in the output schema, be-

cause this information is not available for all annotated tumor entries and multiple documents may be valid when findings are repeated across reports. The manual error analysis is limited to TSIZE errors from three experiments for the Qwen models. We therefore use it to show common error patterns rather than to provide a complete error analysis.

Finally, our setup relies on fixed query templates, fixed chunks, and fixed rules for selecting evidence. It does not adapt retrieval based on intermediate outputs or uncertainty. Future work could explore more adaptive approaches, such as iterative querying, on-demand checking of additional reports, or explicit steps for resolving the status of tumor mentions.

9 Acknowledgments

This work was supported by the Federal Ministry of Research, Technology and Space (BMFTR) through the project **VERANDA** (16KIS2048) and by the Federal Joint Committee (G-BA) though the project **ADBoard** (01VSF21047).

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 1998–2022. Association for Computational Linguistics.

Marco Armbruster, Markus Guba, Joachim Andrassy, Markus Rentsch, Vincent Schwarze, Johannes Rübenthaler, Thomas Knösel, Jens Ricke, and Harald Kramer. 2021. Measuring HCC Tumor Size in MRI—The Sequence Matters! *Diagnostics*, 11(11):2002.

Luc Builtjes and Alessa Hering. 2026. Tracking Cancer Through Text: Longitudinal Extraction From Radiology Reports Using Open-Source Large Language Models. *Preprint*, arXiv:2603.09638.

Victoria Chernyak, Milana Flusberg, Amy Law, Mariya Kobi, Viktoriya Paroder, and Alla M. Rozenblit. 2018. Liver Imaging Reporting and Data System: Discordance Between Computed Tomography and Gadoxetate-Enhanced Magnetic Resonance Imaging for Detection of Hepatocellular Carcinoma Major Features. *Journal of Computer Assisted Tomography*, 42(1):155–161.

Surabhi Datta, Hio Cheng Lam, Atieh Pajouhi, Sunitha Mogalla, and Kirk Roberts. 2022. A Cross-document Coreference Dataset for Longitudinal Tracking across Radiology Reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3686–3695, Marseille, France. European Language Resources Association.

Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*, 11(5):392–402.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Iryna Hartsock, Nikolas Koutsoubis, Sabeen Ahmed, Nathan Parker, Matthew B. Schabath, Cyrillo Araujo, Aliya Qayyum, Cesar Lam, Robert A. Gatenby, and Ghulam Rasool. 2026. Clinical AI in Radiology: Foundations, Trends, Applications, and Emerging Directions. *Cancers*, 18(6):942.

Justin Hofenbitzer, Sebastian Schöning, Belle Sebastian, Jacqueline Lammert, Luise Modersohn, Martin Boeker, and Diego Frassinelli. 2025. GerMedIQ: A Resource for Simulated and Synthesized Anamnesis Interview Responses in German. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, page 1064–1078. Association for Computational Linguistics.

Lisa Kühnel and Juliane Fluck. 2022. We are not ready yet: limitations of state-of-the-art disease named entity recognizers. *Journal of Biomedical Semantics*, 13(1).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Dongchen Li, Jitao Liang, Wei Li, Xiaoyu Wang, Longbing Cao, and Kun Yu. 2026. CliCARE: Grounding Large Language Models in Clinical Guidelines for Decision Support over Longitudinal Cancer Electronic Health Records. *Preprint*, arXiv:2507.22533.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Gabriella Macchia, Gabriella Ferrandina, Stefano Patarnello, Rosa Autorino, Carlotta Masciocchi, Vincenzo Pisapia, Cristina Calvani, Chiara Iacomini,

Alfredo Cesario, Luca Boldrini, Benedetta Gui, Vittoria Rufini, Maria Antonietta Gambacorta, Giovanni Scambia, and Vincenzo Valentini. 2022. Multidisciplinary Tumor Board Smart Virtual Assistant in Locally Advanced Cervical Cancer: A Proof of Concept. *Frontiers in Oncology*, 11.

Daniela Muenzel, Heinz-Peter Engels, Melanie Bruegel, Victoria Kehl, Ernst Rummeny, and Stephan Metz. 2012. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. *Radiology and Oncology*, 46(1).

Joan Okasako and Carolyn Bernstein. 2022. Multidisciplinary tumor boards and guiding patient care: The AP role. *J. Adv. Pract. Oncol.*, 13(3):227–230.

Vivek Patkar, Dionisio Acosta, Tim Davidson, Alison Jones, John Fox, and Mohammad Keshtgar. 2011. Cancer Multidisciplinary Team Meetings: Evidence, Challenges, and the Role of Clinical Decision Support Technology. *International Journal of Breast Cancer*, 2011:1–7.

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420.

Maria Reig, Alejandro Forner, Jordi Rimola, Joana Ferrer-Fàbrega, Marta Burrel, Ángeles Garcia-Criado, Robin K. Kelley, Peter R. Galle, Vincenzo Mazzaferro, Riad Salem, Bruno Sangro, Amit G. Singal, Arndt Vogel, Josep Fuster, Carmen Ayuso, and Jordi Bruix. 2022. BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update. *Journal of Hepatology*, 76(3):681–693.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2026. MedGemma Technical Report. *Preprint*, arXiv:2507.05201.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. *JMIR Medical Informatics*, 12:e55318.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 12739–12753. Association for Computational Linguistics.

Sören Spiegel, Seid Muhie Yimam, Philipp Breitfeld, and Frank Ückert. 2025. Adaption and Evaluation of Generative Large Language Models for German Medical Information Extraction. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 33–47, Hannover, Germany. HsH Applied Academics.

Arndt Vogel, Tim Meyer, Gonzalo Sapisochin, Riad Salem, and Anna Saborowski. 2022. Hepatocellular carcinoma. *The Lancet*, 400(10360):1345–1362.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025a. Medical Graph RAG: Evidence-based Medical Large Language Model via Graph Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 28443–28467. Association for Computational Linguistics.

Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. 2025b. An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications*, 16(1).

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, page 6233–6251. Association for Computational Linguistics.

Wen-wai Yim, Sharon W. Kwan, and Meliha Yetisgen. 2016. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. *Journal of Biomedical Informatics*, 64:179–191.

Imge Yüzüncüoglu, Yuxuan Chen, A. Altar Lüser, Roland Roller, Philippe Thomas, Sebastian Möller, Nikitha Shruthi Ramasetti, Sharlyn ST Ng, Robert Oehring, and Felix Krenzien. 2026. Entwicklung einer automatisierten Informationsextraktion und Entscheidungsunterstützung für die Tumorkonferenz hepatozellulärer Karzinome. In Anne Herrmann, editor, *KI in der Arzt-Patienten-Kommunikation: Ein Weg zu einer menschlicheren Medizin*. Springer Berlin Heidelberg, Berlin, Heidelberg. In press.

Xiaoying Zhong, Siyi Li, Zhao Chen, Long Ge, Dongdong Yu, Shijia Wang, Liangzhen You, and Hongcai Shang. 2025. Considerations for Patient Privacy of Large Language Models in Health Care: Scoping Review. *Journal of Medical Internet Research*, 27:e76571.

A Additional Analyses

A.1 Chronological and type-grouped history

Chronological and type-grouped concatenation perform similarly overall, but type grouping gives small gains in several parameter-level settings. On the test set, type-grouped concatenation matches or exceeds chronological concatenation for TSIZE F1 in four of five models, with improvements of up to 5.8 points. The largest gain appears for Qwen3-32B, where TSIZE F1 increases from 0.742 with chronological history to 0.8 with type-grouped history. For TLOC, the difference is smaller. Type grouping is equal to or better in three of five models, with changes between -0.3 and +0.6 points for most models. The exception is LLaMA3.1-70B, where chronological history is slightly better for both TSIZE and TLOC.

Case-level detection shows even smaller differences. Type grouping and chronological history are tied or nearly tied for several models, and the largest changes are only a few points. This supports the view that report organization mainly affects parameter extraction, where the model must select the correct size and location mention, rather than case-level active tumor detection. One possible explanation is that grouping by document type makes clinically important evidence easier to use together. Radiology reports and discharge letters contain much of the information needed for tumor size, location, and follow-up status. When these reports are grouped, the model may compare similar evidence sources more easily instead of moving through a mixed chronological sequence of radiology, pathology, surgery, and discharge documents. This may explain why type grouping helps TSIZE for some models, especially Qwen3-32B.

However, the effect is not robust enough to treat type grouping as clearly superior. Chronological history preserves the natural temporal order of the patient record, which is important for deciding whether later evidence overrides older findings. Type grouping may make report-type evidence more coherent, but it can also weaken the temporal flow across document types. We therefore treat both variants as full-history strategies in the main analysis. The main finding is that including longitudinal evidence matters more than choosing between chronological and type-grouped ordering.

A.2 Context budget

Figure 4 shows the effect of increasing the context budget for history concatenation. Increasing the context window gives limited and inconsistent gains. For TSIZE, recall remains mostly stable, while precision often decreases as more context is added. This suggests that the additional context introduces older measurements and repeated tumor descriptions without substantially improving evidence coverage. TLOC shows a similar but weaker trend, and case-level performance remains largely unchanged across context budgets.

This pattern is also reflected in the standard deviations in Table 6. For most full-history settings, variability across context budgets is small, typically around $\sigma < 0.01$–0.05. This means that increasing the context window from 8k to 16k or 32k rarely changes performance substantially. The dataset statistics also help explain this result. In the test set, the median case length is 8,368 tokens and the 75th percentile is 12,192 tokens, so many cases already fit mostly within an 8k or 16k budget. Only 11.59% of test cases exceed 16k tokens, which means that very long histories are relatively rare. Overall, larger context windows are not a general solution in this setting. They may help individual long cases, but they also add more outdated or ambiguous evidence that the model has to filter. The main limitation is therefore not only context length, but the ability to select the finding that reflects the active disease state.

A.3 Medical fine-tuning

The main model set includes MedGemma-27B as the only clinically fine-tuned model, so its results cannot separate the effect of medical adaptation from the effect of the underlying model. We therefore additionally evaluate Gemma-3-27B-it, the general-purpose model of the same family and size, under the same prompts, strategies, and context budget. Table 7 reports the comparison.

Analyzing the table along two axes, precision and recall, shows that the two models fail in differently. Gemma-3 has higher TSIZE recall in every setting, by 1.4–5.5 points, but lower TSIZE precision, by 9.7–10.4 points in the history-based settings. Gemma-3 extracts most of the relevant find-

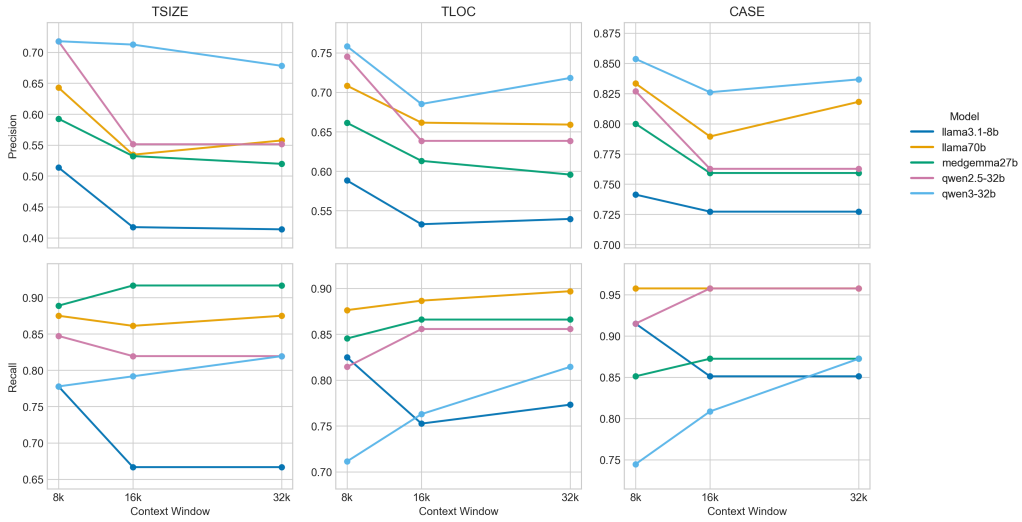

Figure 4: Effect of context budget under history concatenation on test-set precision and recall.

Model	Strategy	TSIZE			TLOC			Case		
		P	R	F1	P	R	F1	P	R	F1
MedGemma-27B	no-history	**0.573**	0.764	**0.655**	**0.632**	0.742	**0.682**	**0.800**	0.851	0.825
	type-grouped concat	**0.593**	0.889	**0.711**	**0.667**	0.845	**0.745**	**0.800**	0.851	0.825
	history concat	**0.593**	0.889	**0.711**	**0.661**	0.845	**0.742**	**0.800**	0.851	0.825
	semantic RAG	**0.438**	0.875	**0.583**	**0.529**	0.845	**0.651**	0.721	0.936	0.815
	time-aware RAG	**0.471**	0.889	0.615	0.534	0.814	0.645	0.717	0.915	0.804
Gemma-3-27B-it	no-history	0.471	**0.792**	0.591	0.547	**0.784**	0.644	0.763	**0.957**	**0.849**
	type-grouped concat	0.496	**0.903**	0.640	0.577	**0.887**	0.699	0.763	**0.957**	**0.849**
	history concat	0.489	**0.903**	0.634	0.581	**0.887**	0.702	0.763	**0.957**	**0.849**
	semantic RAG	0.420	**0.917**	0.576	0.512	**0.856**	0.641	**0.738**	**0.957**	**0.833**
	time-aware RAG	0.456	**0.944**	0.615	**0.538**	**0.876**	**0.667**	**0.726**	**0.957**	**0.826**

Table 7: Test-set comparison of MedGemma-27B-text-it and its general-purpose counterpart Gemma-3-27B-it across context construction strategies. Bold marks the better value of the two models for each metric.

ings but also returns many additional ones. Medical fine-tuning therefore appears to help mainly with the decision of which mentions to leave out, which is the harder part of the task in a longitudinal record.

At the case-level, Gemma-3 is better by 1.8–2.4 points in every setting, driven by a case-level recall of 0.957 against 0.851 for MedGemma. A model that over-extracts is rarely wrong about whether a case contains an active tumor, so the same behavior that lowers parameter precision raises case-level recall. This is consistent with the main results, where case-level detection is less sensitive to context construction than parameter extraction.

Two limitations should be kept in mind. MedGemma is fine-tuned primarily on English clinical text (Sellergren et al., 2026), so this comparison measures the transfer of English medical adaptation to German reports rather than domain adaptation in general. The comparison also covers one model family at one size, so it shows that medical fine-tuning is not sufficient to remove the precision limitation observed for all models, rather than establishing a general effect of clinical pretraining.

A.4 Retrieval ablation

The retrieval pipeline combines sparse and dense candidate retrieval before fusion and reranking (Section 4.3). To test whether both components are needed, we replace the candidate retrieval stage with BM25 alone and with dense retrieval alone, keeping fusion, reranking, temporal selection, and all other hyperparameters fixed. We run the ablation under time-aware RAG with Qwen2.5-32B, the configuration used for hyperparameter tuning, and report it in Table 8. The hybrid row is the configuration used in the main results and is therefore taken directly from Table 6.

The effect of the retriever is small and not uniform across parameters. For TSIZE, all three variants fall within 0.3 points of each other, and the hybrid and dense retrievers are tied. For TLOC and case-level detection the hybrid retriever is best, by 1.1 and 1.7 points over dense retrieval and by 2.1

Retriever	TSIZE F1	TLOC F1	Case F1
BM25 only	0.667	0.709	0.865
Dense only (BGE-M3)	**0.670**	0.719	0.868
Hybrid (BM25 + dense)	**0.670**	**0.730**	**0.885**

Table 8: Retrieval stage ablation under time-aware RAG with Qwen2.5-32B. Only the candidate retrieval stage changes; fusion, reranking, and temporal selection are identical across rows. The hybrid row is the configuration used in the main results and is reported from Table 6.

and 2.0 points over BM25. The gains might be attributed to the cases where the query is a semantic description of an anatomical region rather than a numeric measurement. Tumor sizes are short, surface-level strings that lexical matching already retrieves reliably, so adding dense retrieval contributes little; segment mentions and the surrounding status evidence are phrased more variably, which is where combining lexical and semantic signals helps; however, these differences are small relative to the size of the test set.

A.5 Extended Error Analysis

We manually analyzed TSIZE errors for Qwen models under history concatenation, semantic RAG, and time-aware RAG settings. We categorized errors according to the apparent source of failure: post-treatment interpretation, missing tumor/status evidence, temporal filtering, HCC-status ambiguity, satellite or suspicious lesions, value-level errors, and structured-output errors. Figure 5 shows the resulting distributions.

Active-status interpretation. The largest source of error is post-treatment interpretation, where the model fails to determine whether a previously described lesion should still be considered active after treatment, resection, or follow-up. Averaged across the three context strategies, this category accounts for 52% of TSIZE errors for Qwen2.5-32B and 29.2% for Qwen3-32B, suggesting that explicit reasoning reduces, but does not eliminate, the difficulty. These errors are not simply caused by missing domain knowledge. In several inspected cases, the prompt already specifies relevant decision rules, such as not returning a tumor size after R0/R1 resection or local control when no residual lesion is measured. Nevertheless, the model sometimes identifies the relevant evidence and the applicable rule, but still makes the wrong final active-status decision. This suggests a limitation in reliably applying conditional clinical rules, especially in post-treatment contexts where many treatment types and follow-up formulations cannot be exhaustively enumerated in the prompt. Smaller open-weight models may therefore struggle not only with clinical knowledge, but also with consistently mapping complex treatment evidence to the final extraction decision.

Retrieval-related context gaps. The retrieval-based settings show that temporal selection improves evidence coverage, but does not fully resolve context gaps. Compared with semantic RAG, time-aware RAG reduces missing-status-evidence errors, indicating that temporal structure helps retrieve more relevant follow-up evidence. However, both retrieval variants remain limited by incomplete chunk-level context. In some cases, retrieval selects chunks with older tumor information but fails to retrieve later evidence on treatment response, resection, local control, or residual vitality. In other cases, the retrieved context is too narrow and omits surrounding diagnostic information needed to determine whether a finding corresponds to an HCC tumor. These errors help explain why time-aware RAG improves over semantic RAG, but still remains less precise than full-history input. The lower share of missing-status-evidence errors for Qwen3-32B suggests that reasoning can sometimes compensate for incomplete or noisy retrieved context, although it does not remove the underlying retrieval limitation.

Temporal, suspicious, and output errors. Other errors involve selecting the wrong measurement from the longitudinal record, such as extracting an outdated size, returning both previous and current measurements. For Qwen3-32B under history concatenation, temporal filtering accounts for 17.2% of TSIZE errors. A separate group involves small satellite or HCC-suspicious lesions; these should be interpreted cautiously because some disagreements may reflect annotation uncertainty rather than clear model failure, since the dataset annotates final active findings but not all historical, suspicious, or uncertain lesions. Finally, value-level and structured-output errors include hallucinated or rounded sizes, merged lesions, incomplete entries, and `N/A` values despite available evidence.

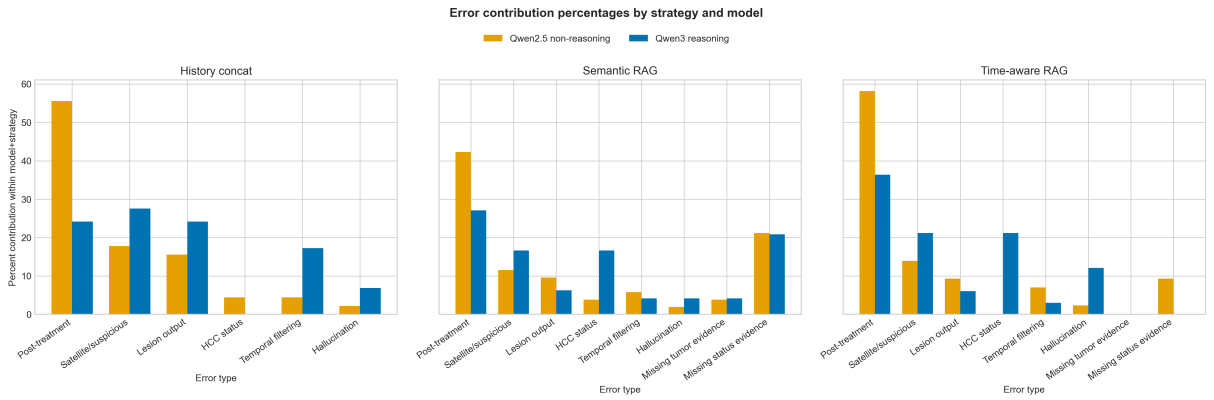

Figure 5: Percentage contribution of error types for Qwen2.5-32B and Qwen3-32B across history concat, semantic RAG, and time-aware RAG settings.

B Hyperparameter Tuning for RAG

We perform hyperparameter tuning for retrieval-based context construction in two stages. First, we optimize chunking and retrieval parameters controlling how much textual evidence is retrieved per report (Table 10). Second, we tune recency weighting parameters to balance temporal relevance and semantic similarity in retrieval (Table 11). All hyperparameter tuning experiments were conducted using Qwen2.5-32B-Instruct in a non-reasoning setup, as it provides strong performance in our setting while requiring significantly less computational cost than reasoning-based variants of the same model family.

In the first stage, we vary chunk size, the target number of tumor-related chunks, and the maximum number of chunks per report. These parameters directly control the trade-off between evidence coverage and noise. Increasing the number of retrieved chunks consistently improves recall, particularly for TSIZE, but introduces more irrelevant context, leading to a drop in precision. Chunk size shows a smaller effect, with 1000 characters providing slightly more stable performance across tasks compared to smaller (800) and larger (1200) chunks.

Based on these observations, we select a configuration of chunk size 1000, target 10 chunks, and a maximum of 3 chunks per report. This setting prioritizes high recall for tumor evidence extraction, which is critical for downstream extraction, while maintaining competitive F1 scores.

We further tune the recency weighting parameters α and λ, which govern the trade-off between semantic similarity and temporal relevance during retrieval. Across all tasks, performance differences remain small, suggesting limited sensitivity to these parameters within the explored range. This likely stems from temporal constraints already imposed during candidate selection, where chunks are sampled across multiple time horizons (e.g., older, intermediate, and recent reports) with a bias toward recent evidence. Consequently, recency is effectively enforced regardless of the scoring function, and α mainly influences the final ranking rather than the temporal diversity of retrieved candidates.

For TSIZE, higher α values (e.g., $\alpha = 0.8$) yield slightly better recall (up to 0.923), suggesting that temporal ranking remains the dominant signal for identifying tumor evidence. Variations in λ have only minor effects, with no consistent trend across configurations. For TLOC, performance is even more stable, with F1 scores clustered tightly around 0.72–0.73, and only marginal precision gains observed for lower α values. Case-level classification is largely unaffected by these parameters, with near-constant recall and minimal variation in F1.

Based on these observations, we select $\alpha = 0.8$ and $\lambda = 0.001$, as this configuration provides strong and stable performance across tasks while preserving high recall for tumor-specific attributes without introducing unnecessary noise.

C Prompt Templates

The prompts were written in German, matching the language of the clinical reports. All experiments use the German versions as shown in Appendix C.1 and Appendix C.2. For translation purposes, an English gloss for prompts is given in Appendix C.3. The gloss is machine-translated and is provided for illustration only; it was never used in any experiment.

C.1 Direct History Prompt Template

```
<|im_start|>system
Antworte ausschließlich mit gültigem JSON. Gib keinen Erklärtext, keine Kommentare und keine Markdown-Formatierung aus. Die Antwort muss direkt mit '{' beginnen und mit '}' enden.

Analysiere medizinische Texte und extrahiere HCC-Tumore (hepatozelluläres Karzinom) oder HCC-typische bzw. HCC-verdächtige Läsionen (z.B. LI-RADS 4/5, arterielle Hypervaskularisation, Washout, Diffusionsrestriktion, EOB-Defekt). Dazu gehören auch Läsionen mit typischen radiologischen Merkmalen oder Läsionen, die explizit als HCC-typisch bzw. HCC-suspekt bezeichnet werden.

Extrahiere Läsionen, die im neuesten relevanten Verlauf weiterhin als aktiv vorhanden oder radiologisch vital nachweisbar sind. Behandelte Läsionen sollen nur dann extrahiert werden, wenn trotz Therapie weiterhin Vitalitätszeichen bestehen.

Wenn ein Tumor in älteren Befunden beschrieben wird und in einem neueren Befund explizit als entfernt, behandelt ohne Residuum oder rezidivfrei beschrieben wird, darf er nicht extrahiert werden.

Läsionen nach Therapie (z.B. TACE, Brachytherapie, Ablation, Resektion) gelten NICHT als aktive HCC-Tumore, wenn neuere Befunde sie als gut kontrolliert, lokal kontrolliert, devaskularisiert, nekrotisch oder als rein posttherapeutische Reststruktur ohne Vitalitätszeichen beschreiben (z.B. Narbe, Ablationszone oder fehlendes Enhancement).

Morphologisch persistierende Läsionen nach Therapie dürfen nur dann extrahiert werden, wenn weiterhin eindeutige Vitalitätszeichen bestehen, z.B. arterielle Hypervaskularisation, Washout, Diffusionsrestriktion, noduläre kontrastmittelaufnehmende Anteile, explizit vitales Residuum oder Progress.

Pathologieberichte dienen primär zur Bestätigung von Histologie und Tumorgröße im resezierten oder behandelten Material, jedoch in der Regel NICHT zur Beurteilung aktuell aktiver intrahepatischer Tumoren. Für die Aktivitätsbewertung haben neuere radiologische oder klinische Verlaufskontrollen Vorrang.

Eine Läsion darf anhand eines Pathologieberichts nur dann als aktuell aktiv gewertet werden, wenn explizit ein vitales Residuum, ein fortbestehender nicht entfernter Tumor oder eine unvollständige Entfernung mit makroskopischem Residuum (R2) beschrieben wird; eine R1-Resektion (mikroskopisches Residuum) gilt nicht als messbare Läsion, daher ist "tumor_size": "N/A" zu setzen, während die Tumorlokalisation weiterhin extrahiert werden soll, sofern sie eindeutig bestimmbar ist, ohne frühere Größen zu übernehmen; eine R2-Resektion (makroskopisches Residuum) gilt als aktiver Tumor, wobei Tumorgröße und Lokalisation nur dann extrahiert werden dürfen, wenn sie im aktuellen Befund explizit angegeben sind.
<|im_end|>
<|im_start|>user
Analysiere die folgenden medizinischen Befunde.

Die Befunde enthalten Datumsangaben in den Überschriften. Verwende die Datumsangaben, um den zeitlichen Verlauf zu verstehen. Neuere Befunde überschreiben ältere Befunde bei widersprüchlichen Aussagen. Nutze die neuesten relevanten Informationen für die finale Entscheidung.

Extrahiere alle HCC- oder HCC-suspekten Läsionen, die nach Berücksichtigung des Verlaufs noch aktuell aktiv relevant sind.

Wenn ein Befund eine aktuelle Größe und eine frühere Größe (z. B. mit "VU") nennt, extrahiere nur die aktuelle Größe.

Wenn eine Information fehlt, verwende "N/A".

Wenn keine passenden Läsionen gefunden werden, gib eine leere Liste zurück.

WICHTIGE REGELN
* Ein Satz kann mehrere Tumore enthalten. Jeder Tumor muss als eigener Eintrag erscheinen.
* Die anatomische Lage (z.B. Segment) kann vor oder nach der Größe stehen und muss korrekt zugeordnet werden.
* Tumorgrößen können 1D, 2D oder 3D angegeben sein. Verwende die vollständigste vorhandene Größenangabe.
* Wenn dieselbe Läsion mit gleicher Größe und Lage mehrfach erwähnt wird, darf sie nur einmal im JSON erscheinen.
* Wenn mehrere Tumore im gleichen Segment mit unterschiedlichen Größen beschrieben werden, müssen sie als separate Einträge erscheinen.
* Pathologieberichte sollen primär für Histologie, Größe und Resektionsstatus genutzt werden und nur bei explizitem vitalem Residuum, Tumor am Rand oder unvollständiger Entfernung zur Aktivitätsbewertung beitragen.

BEISPIEL-FORMAT

{
"active_hcc_tumors": [
{
"tumor_size": "<Größe mit Einheit oder N/A>",
"tumor_location": "<Segment oder andere anatomische Lage oder N/A>",
"source_date": "<Datum des verwendeten Befundes oder N/A>",
"source_doc_type": "<Dokumenttyp oder N/A>"
}
]
}

BEFUNDE:
Relevante medizinische Befunde zu Tumoren:
----- 2024-02-24 Arztbrief -----
<full report>>

----- 2024-03-10 Bildgebung -----
<full report>>
....
<|im_end|>
<|im_start|>assistant
```

C.2 RAG Prompt Template

```
<|im_start|>system
The same system prompt as above.
<|im_end|>
<|im_start|>user
The same user prompt as above until this part.

BEFUNDE:
Relevante medizinische Befunde zu Tumoren (älteste zuerst):
----- 2024-02-24 Arztbrief -----
<chunk>
----- 2024-03-10 Bildgebung -----
<chunk>

Informationen zu möglichen Resektionen (älteste zuerst):
----- 2024-02-24 Arztbrief -----
<chunk>
 ----- 2024-02-24 OP-Berichte -----
<chunk>
<|im_end|>
<|im_start|>assistant
```

C.3 English Gloss of Prompts

Gloss of the direct-history prompt (Appendix C.1). The RAG prompt (Appendix C.2) is identical except that the evidence block contains retrieved chunks, grouped into tumor-related and resection-related evidence and ordered oldest first, instead of full reports.

```
<|im_start|>system
Respond exclusively with valid JSON. Do not output explanatory text, comments, or Markdown formatting. The answer must begin directly with '{' and end with '}'.
```

```
Analyze medical texts and extract HCC tumors (hepatocellular
      carcinoma) or HCC-typical or HCC-suspicious lesions (e.g.
       LI-RADS 4/5, arterial hypervascularization, washout,
      diffusion restriction, EOB defect). This also covers
      lesions with typical radiological features, or lesions
      explicitly described as HCC-typical or HCC-suspicious.

Extract lesions that are still present as active, or
      radiologically detectable as viable, in the most recent
      relevant follow-up. Treated lesions are to be extracted
      only if signs of viability persist despite therapy.

If a tumor is described in older reports and a more recent
      report explicitly describes it as removed, treated
      without residue, or recurrence-free, it must not be
      extracted.

Lesions after therapy (e.g. TACE, brachytherapy, ablation,
      resection) do NOT count as active HCC tumors if more
      recent reports describe them as well controlled, locally
       controlled, devascularized, necrotic, or as purely post-
      therapeutic residual structure without signs of
      viability (e.g. scar, ablation zone, or absent
      enhancement).

Morphologically persistent lesions after therapy may be
      extracted only if unambiguous signs of viability remain,
       e.g. arterial hypervascularization, washout, diffusion
      restriction, nodular contrast-enhancing components,
      explicitly viable residue, or progression.

Pathology reports serve primarily to confirm histology and
      tumor size in the resected or treated material, but
      generally NOT to assess currently active intrahepatic
      tumors. For activity assessment, more recent
      radiological or clinical follow-up takes precedence.

A lesion may be counted as currently active on the basis of a
      pathology report only if a viable residue, a persisting
      non-removed tumor, or an incomplete removal with
      macroscopic residue (R2) is explicitly described; an R1
      resection (microscopic residue) does not count as a
      measurable lesion, so "tumor_size": "N/A" is to be set,
      while the tumor location should still be extracted where
       unambiguously determinable, without carrying over
      earlier sizes; an R2 resection (macroscopic residue)
      counts as an active tumor, where tumor size and location
       may be extracted only if explicitly stated in the
      current report.
<|im_end|>
<|im_start|>user
Analyze the following medical reports.

The reports carry dates in their headings. Use the dates to
      understand the temporal course. More recent reports
      override older reports where statements conflict. Use
      the most recent relevant information for the final
      decision.

Extract all HCC or HCC-suspicious lesions that are still
      currently active and relevant once the course has been
      taken into account.

If a report gives a current size and an earlier size (e.g.
      marked "VU", prior examination), extract only the
      current size.

If a piece of information is missing, use "N/A".

If no matching lesions are found, return an empty list.

IMPORTANT RULES

* One sentence may contain several tumors. Each tumor must
      appear as its own entry.
* The anatomical location (e.g. segment) may precede or follow
       the size and must be assigned correctly.
* Tumor sizes may be given in 1D, 2D, or 3D. Use the most
      complete size specification available.
* If the same lesion is mentioned repeatedly with the same
      size and location, it may appear only once in the JSON.
* If several tumors in the same segment are described with
      different sizes, they must appear as separate entries.
* Pathology reports are to be used primarily for histology,
      size, and resection status, and contribute to the
      activity assessment only in the case of explicit viable
      residue, tumor at the margin, or incomplete removal.

EXAMPLE FORMAT

{
"active_hcc_tumors": [
{
"tumor_size": "<size with unit or N/A>",
"tumor_location": "<segment or other anatomical location or N/
      A>",
"source_date": "<date of the report used or N/A>",
"source_doc_type": "<document type or N/A>"
}
]
}

REPORTS:
Relevant medical reports on tumors:
----- 2024-02-24 Physician letter -----
<full report>
....
<|im_end|>
<|im_start|>assistant
```

D Additional Tables

Model	Strategy	Context	TSIZE				TLOC				Case			
			P	R	F1	σ	P	R	F1	σ	P	R	F1	σ
LLaMA-3.1-8B	no-history	16384	0.565	0.765	0.650	0.031	0.500	0.735	0.595	0.067	0.763	0.879	0.817	0.069
	type-grouped concat	32768	0.566	**0.922**	**0.701**	0.010	**0.588**	**0.959**	**0.729**	0.014	0.780	**0.970**	**0.865**	0.009
	history concat	32768	**0.592**	0.824	0.689	0.020	0.556	0.816	0.661	0.018	0.769	0.909	0.833	0.003
	semantic RAG	8192	0.404	0.824	0.542	–	0.419	0.735	0.533	–	0.784	0.879	0.829	–
	time-aware RAG	8192	0.483	0.824	0.609	–	0.494	0.837	0.621	–	**0.806**	0.879	0.841	–
MedGemma-27B	no-history	8192	**0.695**	0.804	0.745	0.008	0.655	0.735	0.692	0.005	0.824	0.848	0.836	0.009
	type-grouped concat	16384	0.676	0.980	**0.800**	0.030	**0.676**	**0.939**	**0.786**	0.018	**0.825**	**1.000**	**0.904**	0.027
	history concat	16384	0.644	0.922	0.758	0.015	0.672	0.878	0.761	0.007	0.821	0.970	0.889	0.018
	semantic RAG	8192	0.532	0.980	0.690	–	0.603	0.898	0.721	–	0.786	**1.000**	0.880	–
	time-aware RAG	8192	0.593	**1.000**	0.745	–	0.603	0.898	0.721	–	0.780	0.970	0.865	–
Qwen2.5-32B	no-history	16384	0.732	0.804	0.766	0.023	0.696	0.653	0.674	0.033	**0.867**	0.788	0.825	0.009
	type-grouped concat	16384	**0.817**	0.961	**0.883**	0.041	**0.754**	0.878	**0.811**	0.011	0.842	**0.970**	**0.901**	0.013
	history concat	16384	0.800	0.941	0.865	0.035	0.750	0.857	0.800	0.009	0.861	0.939	0.899	0.011
	semantic RAG	8192	0.658	**0.980**	0.787	–	0.667	**0.898**	0.765	–	0.780	**0.970**	0.865	–
	time-aware RAG	8192	0.746	**0.980**	0.847	–	0.733	**0.898**	0.807	–	0.842	**0.970**	**0.901**	–
Qwen3-32B	no-history	16384	0.907	0.765	0.830	0.006	0.814	0.714	0.761	0.024	0.933	0.848	0.889	0.043
	type-grouped concat	32768	**0.940**	**0.922**	**0.931**	0.055	**0.891**	**0.837**	**0.863**	0.043	**0.938**	**0.909**	**0.923**	0.036
	history concat	16384	**0.940**	**0.922**	**0.931**	0.037	0.833	0.816	0.825	0.021	**0.938**	**0.909**	**0.923**	0.030
	semantic RAG	8192	0.758	**0.922**	0.832	–	0.690	0.816	0.748	–	0.811	**0.909**	0.857	–
	time-aware RAG	8192	0.839	**0.922**	0.879	–	0.769	0.816	0.792	–	0.882	**0.909**	0.896	–
LLaMA-3.1-70B	no-history	16384	**0.689**	0.824	0.750	0.017	0.610	0.735	0.667	0.009	0.857	0.909	0.882	0.002
	type-grouped concat	8192	0.652	0.882	0.750	0.015	0.641	0.837	0.726	0.003	0.838	0.939	0.886	0.005
	history concat	8192	0.682	0.882	**0.769**	0.034	**0.662**	0.878	**0.754**	0.019	**0.865**	**0.970**	**0.914**	0.019
	semantic RAG	8192	0.481	**1.000**	0.650	–	0.550	**0.898**	0.682	–	0.780	**0.970**	0.865	–
	time-aware RAG	8192	0.515	0.980	0.676	–	0.544	0.878	0.672	–	0.795	0.939	0.861	–

Table 9: Dev-set performance of context construction strategies across models. Models are ordered by model size. Bold marks the best value within each model family for each metric; ties are bolded jointly. σ denotes variability across evaluated context lengths and is reported only for strategies tested with multiple context budgets.

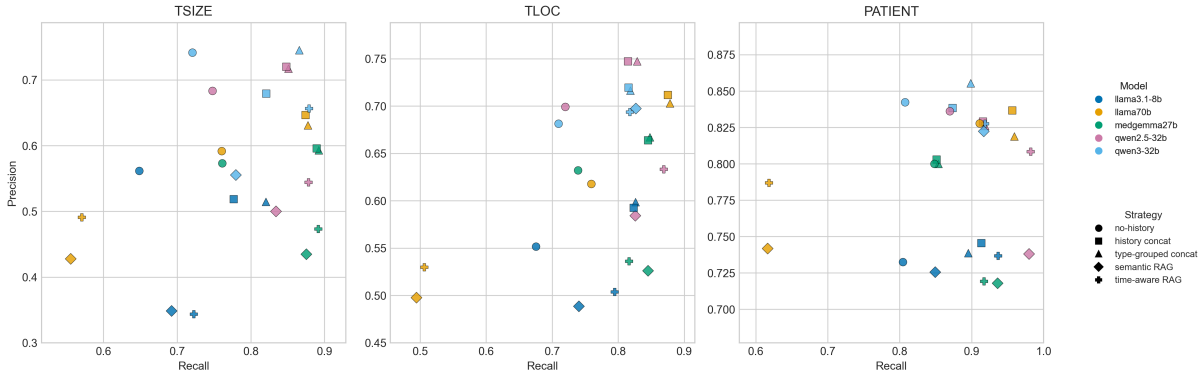

Figure 6: Zoomed version of test-set precision–recall trade-offs across strategies for TSIZE, TLOC, and case-level tumor detection.

Chunk	Target	Max	TSIZE			TLOC			Case		
			P	R	F1	P	R	F1	P	R	F1
1000	10	3	0.594	0.891	**0.713**	0.634	0.865	0.731	0.673	0.972	0.795
1000	8	2	0.561	0.859	0.679	**0.643**	0.865	**0.738**	0.673	0.972	0.795
1000	6	2	0.589	0.828	0.688	0.642	0.823	0.721	**0.706**	**1.000**	**0.828**
1200	10	3	0.567	0.859	0.683	0.619	**0.896**	0.732	0.636	0.972	0.769
1200	8	2	0.547	**0.906**	0.682	0.594	0.885	0.711	0.655	**1.000**	0.791
1200	6	2	0.564	0.828	0.671	0.636	0.854	0.729	0.673	0.972	0.795
800	10	3	0.533	0.875	0.663	0.591	0.844	0.695	0.643	**1.000**	0.783
800	8	2	0.553	0.891	0.683	0.571	0.833	0.678	0.660	0.972	0.787
800	6	2	0.570	0.828	0.675	0.639	0.812	0.716	0.679	**1.000**	0.809

Table 10: Hyperparameter sweep for chunk size and the number of tumor information chunks along with how many chunks a report can contribute at max. Increasing the number of retrieved chunks (target/max) improves recall but consistently reduces precision, especially for TSIZE. The selected configuration (1000, 10, 3) maximizes TSIZE recall and overall F1, at the cost of lower precision.

α	λ	TSIZE			TLOC			Case		
		P	R	F1	P	R	F1	P	R	F1
0.8	0.0075	0.545	**0.923**	**0.686**	0.634	0.856	0.728	0.780	0.979	0.868
0.8	0.001	0.536	0.908	0.674	0.627	0.866	0.727	**0.807**	0.979	**0.885**
0.8	0.003	0.531	**0.923**	0.674	0.632	0.866	0.730	0.780	0.979	0.868
0.65	0.0075	0.526	**0.923**	0.670	0.627	0.866	0.727	0.780	0.979	0.868
0.65	0.001	0.532	0.892	0.667	0.625	**0.876**	0.730	0.793	0.979	0.876
0.65	0.003	0.518	0.908	0.659	0.618	0.866	0.721	0.793	0.979	0.876
0.5	0.001	0.527	0.892	0.663	**0.636**	0.866	**0.734**	0.793	0.979	0.876
0.5	0.003	0.522	0.908	0.663	0.620	**0.876**	0.726	0.793	0.979	0.876
0.5	0.0075	0.513	0.908	0.656	0.618	0.866	0.721	0.793	0.979	0.876

Table 11: Hyperparameter sweep for recency weighting. α controls the balance between semantic similarity and temporal relevance, while λ determines the decay applied to older documents. Performance differences are small across configurations, with slightly higher recall observed for larger α values and minor precision gains for lower α.

Hyperparameter	**Value**
Chunking	
Chunk size	1000 characters
Sentence overlap	2 sentences
Retrieval	
Tumor initial retrieval size (n)	80
Resection initial retrieval size (n)	30
BM25 enabled	True
BM25 tumor n	80
BM25 resection n	30
BM25 minimum score	0.1
RRF constant (k)	60
Tumor candidate pool (k)	180
Resection candidate pool (k)	60
Time-aware scoring	
Recency enabled	True
Tumor recency weight (α_{tumor})	0.8
Tumor recency decay (λ_{tumor})	0.001
Temporal constraints	
Tumor bucket minima (old, middle, recent)	(1, 1, 3)
Latest-report guarantee (tumor)	True
Latest-chunk guarantee (resection)	True
Max tumor chunks per report	3
Max resection chunks per report	2
Final context	
Selected tumor chunks	10
Selected resection chunks	3
Embeddings and reranking	
Embedding model	`BAAI/bge-m3`
Cross-encoder model	`BAAI/bge-reranker-v2-m3`
Reranker batch size	16
Keep top-n after rerank	0 (disabled)
LLM inference	
Model	Table 5
Max output tokens	4096
Max model context length	32768
Temperature	0.1
Robustness	
Summary retry limit	3
Extra JSON retries	3
Query templates	
Tumor Template 1	`HCC hepatozelluläres Karzinom Leber Läsion Herd Herdbefund Raumforderung Segment Seg. S1 S2 S3 S4 S5 S6 S7 S8 Tumordurchmesser Durchmesser Größe mm cm`
Tumor Template 2	`LI-RADS HCC-typisch arterielle Hypervaskularisation Washout Diffusionsrestriktion EOB-Defekt`
Tumor Template 3	`Leberläsion Verlauf VU Größenzunahme mehrere Läsionen Segment Seg. S1 S2 S3 S4 S5 S6 S7 S8 Tumordurchmesser Durchmesser`
Resection Template 1	`Resektion Teilresektion Hemihepatektomie R0 R1 R2 Absetzungsrand Resektionsrand`
Resection Template 2	`Ablation RFA MWA TACE SIRT postinterventionell behandelte Läsion Nekrose Ablationszone`
Resection Template 3	`kein Residuum kein Tumornachweis rezidivfrei nicht mehr nachweisbar ohne Restvitalität kein vitaler Tumor`

Table 12: Full configuration used for all retrieval-based experiments, including the query templates used for candidate retrieval.