Generating Linguistically Diverse and Semantically Faithful Patient Reports

Alejandra Lorenzo[1,2] **Adrien Coulet**[3] **Claire Gardent**[1]
[1]CNRS, Université de Lorraine, Loria UMR7503, Nancy, France
[2]U2R, Unité de Radiothérapie République, Clermont-Ferrand, France
[3]Inria, Inserm, Université Paris Cité, HeKA UMR1346, Paris, France
Alejandra.Abdelnur@loria.fr Adrien.Coulet@inria.fr Claire.Gardent@loria.fr
Correspondence: Alejandra.Abdelnur@loria.fr

Abstract

To protect patient privacy, medical models are frequently trained on synthetic datasets. Focusing on 13 clinical attributes commonly found in breast cancer records, we propose a synthetic data generation framework in which each generated sample is a data–report pair: the structured data capture the 13 attributes of interest, while the accompanying (French) text simulates a corresponding breast cancer pathology report. Data privacy and coverage are obtained by generating the structured data using a tabular data synthesiser. In this work, we focus on enhancing the diversity and quality of the generated texts through a combination of data-driven knowledge distillation, data filtering and data combination. Our findings suggest that the quality of the training data used for the student model outweighs its quantity; that combining data from multiple teachers helps improve performance; and that a compact model (Mistral 7B) can match the zero-shot Information Extraction performance of a significantly larger model (MedGemma 27B) when fine-tuned on high-quality synthetic data[1].

1 Introduction

Processing medical data raises important challenges for privacy, governance, and reproducibility. Clinical datasets contain sensitive personal information and must therefore be handled within secure infrastructures, under appropriate pseudonymisation and regulatory oversight. These constraints limit data sharing and complicate the reproducibility of computational studies.

Synthetic data provide an alternative by enabling experimentation without direct access to protected patient records. In this work, we focus on generating synthetic data–report pairs for information extraction from French breast-cancer pathology reports. Each dataset instance consists of a data–report pair, where the data is a set of 13 clinically relevant attributes and their values. The associated texts simulate patient reports and, as such, should verbalise the data faithfully while preserving the linguistic and structural variability observed across physicians. Figure 1 shows an example.

Building on Lorenzo et al. (2026), we first generate differentially private structured records using PATE-CTGAN. This yields a dataset of synthetic structured data that preserves patient privacy while effectively simulating real patient data.

We then focus on generating synthetic patient reports from this data seeking to optimise two key criteria: faithfulness and diversity. Faithfulness requires that the generated reports verbalise all and only the attribute–value pairs present in the input to generation. Diversity refers to linguistic diversity. We strive to generate reports that capture variability in report writing, the fact that physicians each have their own style and way of writing.

To optimise these two dimensions, we explore different ways of creating the data-text corpus. We compare four teacher configurations based on two language-model families, with and without retrieval from instantiated clinician templates provided with the original data – these templates provide examples of how to verbalise a set of attribute–value pairs and are instantiated using clinically plausible values. We explore the trade-off between faithfulness filtering, which decreases the amount of data available but improves its quality, and data quantity. And we study how combining data produced by different models impacts results.

When evaluating, we assess both the quality of the generated synthetic data, in terms of diversity and faithfulness, and its effectiveness in supporting the training of information extraction models that can extract the 13 clinical attributes and their values from patient reports. Our results show that

[1]A repository containing the code of this work is available at https://github.com/lorenzoale/Quality_Synthetic_Pathological_Reports

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 24–48
November 6-10, 2026 ©2026 Association for Computational Linguistics

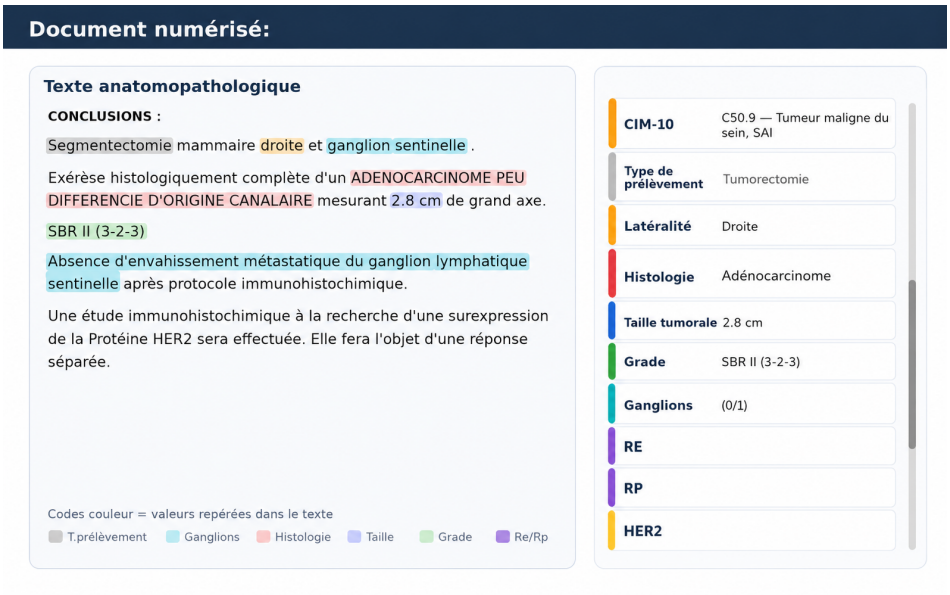


Figure 1: **Example Data-Report Pair.** The diagnosis form (Right) contains some of the attributes discussed in this work. The text on the left shows the corresponding pathology report fragment.

smaller datasets filtered for faithfulness can outperform larger, unfiltered datasets, and that selectively combining complementary teacher outputs can further improve performance.

For downstream evaluation, we fine-tune LLM-based Information Extraction models on our best-performing synthetic datasets and evaluate them on 377 real reports that were manually annotated with the corresponding structured data by domain experts. We find that Mistral-7B improves its F1 score from 0.57 to 0.86, matching the performance of zero-shot MedGemma-27B (0.86), while remaining below fine-tuned MedGemma-27B, which achieves an F1 score of 0.90.

Our contributions can be summarised as follows:

- Building on privacy-preserving tabular synthesis, we study how teacher choice, retrieval-guided generation, faithfulness-based filtering, and selective dataset combination affect synthetic pathology report generation.
- We show that filtered datasets can outperform larger unfiltered datasets, and that selectively combining complementary teacher-generated datasets can further improve faithfulness and downstream utility.
- We validate the LLM-based faithfulness evaluator against clinical annotators and assess downstream information-extraction performance on 377 real, expert-annotated reports.

2 Context

The clinical data we aim to simulate (see Figure 1 for an example) were collected from four clinics located in different regions of France and belonging to the U2R group. They consist of 3,880 structured breast cancer patient records. Patient information handling was conducted in accordance with the GDPR and French regulatory requirements.

The patient *reports* are written in French and exhibit substantial variability in both style and structure, making it essential to generate synthetic texts that reflect that diversity faithfully (Munzone et al., 2024; Grothey et al., 2025; Park et al., 2025).

The *data* we seek to extract from these reports consists of the thirteen clinically relevant attributes listed in Table 1. See Appendix A.1 for more details on how these attributes are discretized and handled to generate synthetic tabular data.

3 Related Work

Synthetic Data Generation for the Medical Domain. Hiebel et al. (2023) show that fine-tuning autoregressive language models on synthetic NER data helps improve clinical NER performance. In a similar direction, Šuvalov et al. (2025) leverage large language models to generate and annotate synthetic health care data for NER. The synthetic data is then used to train a high-performing NER model that can be applied to real-world medical texts while preserving patient privacy. Different from these works, we condition generation on syn-

Attribute	Type	Description / Possible values
Test used for diagnosis	Enumerate	lumpectomy; biopsy; mastectomy; surgical excision; lymph node dissection
Morphology	Enumerate	Morphology code according to the International Classification of Diseases for Oncology (ICD-O), French version
SBR grade	Enumerate	Scarff-Bloom-Richardson (SBR) grade: SBR 1; SBR 2; SBR 3
Tumor size	Numeric	Tumor size (usually expressed in millimeters or centimeters)
Clear margins	Enumerate	yes; no. Indicates whether healthy tissue surrounds the excised tumor
Number of lymph nodes taken	Numeric	Total number of lymph nodes removed during surgery
Number of lymph nodes affected	Numeric	Number of lymph nodes showing tumor involvement
Presence of lymphovascular invasion	Enumerate	yes; no. Indicates tumor invasion of lymphatic or blood vessels
Presence of extracapsular tumor spread	Enumerate	yes; no. Indicates tumor extension beyond the lymph node capsule
Estrogen receptor status	Enumerate	positive; negative
Progesterone receptor status	Enumerate	positive; negative
HER2 status	Enumerate	Human epidermal growth factor receptor 2: positive; negative
Ki-67 percentage	Numeric	Percentage of Ki-67 nuclear protein expression

Table 1: Pathological and molecular attributes extracted from the pathology report

thetic tabular data and explore data driven knowledge distillation, filtering and combination to generate synthetic texts.

Differentially Private Synthetic Text Generation Recent work has explored the generation of synthetic text under differential privacy. Yue et al. (2023) propose a practical framework for generating synthetic textual data while providing differential privacy guarantees, whereas Xie et al. (2024) study differentially private synthetic text generation using foundation model APIs. These approaches address settings in which privacy protection is applied directly to sensitive textual data or to the process used to generate synthetic text from them. Our setting differs in where differential privacy is introduced in the pipeline. We first generate differentially private structured records using PATE-CTGAN, and the language models used for report generation operate only on this synthetic structured data. Also, our contribution is not a new text level DP mechanism, but an investigation of how teacher source, retrieval, faithfulness-based filtering, and selective combination affect synthetic report quality and downstream information extraction utility.

Synthetic Pathology Reports and Data-to-Text Generation Pathology reports contain rich diagnostic information, but their use for model training is limited by privacy constraints and annotation cost. Most prior work has focused on the inverse task: extracting structured data from free-text pathology reports, using rule-based methods, traditional NLP, BERT-based models, and, more recently, LLMs (Nguyen et al., 2012; Munzone et al., 2024; Grothey et al., 2025; Shahid et al., 2025; Park et al., 2025). Our task can also be framed as a domain-specific data-to-text generation problem, where structured clinical attributes are verbalized into coherent report text. Prior data-to-text work has considered inputs such as meaning representations, AMRs, knowledge graphs, and ta-

bles (Gardent et al., 2017; Liu et al., 2018; Parikh et al., 2020; Song et al., 2025). However, pathology report generation differs from general domain data-to-text generation because the output must preserve clinically meaningful variables, follow domain specific reporting conventions, and avoid unsupported findings. Related work by Hackl et al. (2025) develops a synthetic oncology pathology dataset for evaluating LLMs in medical text classification, showing the value of synthetic pathology text as a privacy-preserving resource. Closest to our work, Lorenzo et al. (2026) generate synthetic breast-cancer pathology reports from structured attribute–value pairs and use them to train an information-extraction model. Building on this framework, we study how teacher choice, retrieval-guided generation, faithfulness-based filtering, and selective dataset combination affect report quality and downstream utility.

Knowledge Distillation and Retrieval-Guided Generation. Knowledge distillation enables compact student models to learn from stronger teacher models (Hinton et al., 2015). In natural language generation, distillation has been applied to machine translation, summarization, dialogue, and other generation tasks, often through sequence-level distillation or pseudo-target training (Kim and Rush, 2016; Calderon et al., 2023). Recent LLM-based approaches use teacher-generated outputs as synthetic supervision for smaller models, including in clinical NLP settings where annotated data are scarce (Woo et al., 2025; Lee et al., 2025). We adapt this idea to pathology report generation by using data-driven knowledge distillation to improve text generation from structured diagnostic data. In addition, we explore retrieval-augmented generation with pre-instantiated pathology templates as in-context examples. These templates guide the model toward clinically appropriate structure and terminology, while the structured data determine the factual content.

4 Method

We aim to generate a synthetic dataset (data-report pairs) that supports information extraction.

4.1 Generating Synthetic Tabular Data

Following Lorenzo et al. (2026), we start by generating synthetic data that resembles the original patient data while preserving patient privacy. We derive the synthetic structured data from the 3,880 patient records provided by the U2R group (cf. Section 2) using the PATE-CTGAN tabular data synthesiser (Xu et al., 2019). To reduce the risk of linkage attacks based on unique numerical patterns and to stabilize synthetic data generation (Ganev et al., 2025), we first discretize the data. We then use the PATE-CTGAN tabular data synthesiser to generate two non-overlapping datasets, each containing the 13 clinical attributes of interest: AV_TEACHER (1,062 records) and AV_STUDENT (1,061 records). We use AV_TEACHER as an unlabeled input corpus for teacher inference: each instance in AV_TEACHER is passed to the teacher models, which generate the target outputs used for synthetic supervision. The student model is then trained on each teacher-labeled version of AV_TEACHER. The second dataset, AV_STUDENT, is held out and used to evaluate the performance of the trained student model. Details on the discretization process as well as evaluations of privacy preservation and similarity to the original data are provided in Appendix A. Based on the privacy–utility analysis in Appendix A.7, we retain $\epsilon = 3$ for PATE-CTGAN. All subsequent experiments use structured synthetic records generated under this setting.

4.2 Generating Synthetic Patient Reports

In this paper, we focus on how to generate faithful and diverse reports from these synthetic structured records. Because the available real patient reports contain sensitive information, they cannot directly serve as supervision data for report generation. We therefore use teacher-generated reports as synthetic supervision in a data-driven knowledge distillation (DDKD) framework. Unlike classical knowledge distillation, our primary objective is not model compression. Instead, we use a fixed student architecture to study how the source and curation of teacher-generated supervision affect the resulting report generator. The supervision may differ through the teacher model, retrieval-guided generation, faithfulness-based filtering, or the combination of outputs from multiple teachers. We investigate how these different dimensions affect faithfulness, linguistic diversity and downstream utility. Our hypothesis is that combining complementary teacher-generated datasets may increase diversity, while filtering and selective combination may improve faithfulness by reducing erroneous or inconsistent supervision.

We evaluate the approach both intrinsically and extrinsically. Intrinsic evaluation measures report

faithfulness and linguistic diversity, while extrinsic evaluation assesses downstream information extraction by comparing base models with versions fine-tuned on the synthetic datasets.

5 Evaluation

To evaluate the datasets generated by the different models, we use LLM-as-a-Judge for faithfulness and 9 existing metrics for lexical diversity and richness. We also evaluate the downstream use of these datasets for information extraction by comparing a base IE model (MEDGEMMA and Mistral) with the same model fine tuned on the best synthetic datasets.

5.1 Evaluating Faithfulness

As no reference texts are available, standard reference-based metrics such as BLEU or ROUGE cannot be used. Following Kasner and Dušek (2024), we therefore adopt an LLM-as-a-Judge evaluation protocol, where generated reports are evaluated by prompting an LLM to check whether the synthetic text correctly mentions the input attribute–value pairs. For each input attribute–value pair, the LLM judge is prompted to output a JSON object with two fields. The field '`correct` (yes/no)' indicates whether the text correctly mentions the input attribute–value. The field '`reference`' provides the supporting text span when applicable. We experiment with two LLMs (GPT-5.1 and Gemini-2.5-Pro) as LLM-judge.

We evaluate the reliability of these two LLM judges by comparing their predictions against human judgments. Two clinical experts were asked to manually annotate 50 synthetic pathology reports and their associated data, indicating for each input attribute–value pair whether it was correctly expressed by the text. We then computed the Inter Annotator Agreement (IAA) between the two human annotators and the agreement between each LLM judge and the human annotators. Cohen's κ between the two human annotators was 0.90, indicating almost perfect agreement (Cohen, 1960).

For Humans–LLM agreement, we use Krippendorff's α (Krippendorff, 2011), which is suitable for multiple annotators ($n = 3$ in our case). Both LLMs show high agreement with the human annotators, with α scores of 0.90 for GPT 5.1 and 0.82 for Gemini-2.5-Pro. Based on these results, we used GPT 5.1 for the faithfulness evaluation. The prompts used for the LLM judges and the annotation instructions provided to the human evaluators are reported in Appendix C.2.

We define *faithfulness accuracy* as the proportion of input attribute–value pairs that are correctly expressed in a generated report according to the LLM judge.

5.2 Evaluating Diversity

We compare synthetic and original reports using document level metrics that capture lexical volume, lexical diversity, structural properties, repetitiveness, and fluency. Lexical volume is measured using text length and the number of unique tokens. Lexical diversity is assessed using type-token ratio (TTR) and measure of textual lexical diversity (MTLD). Structural properties are captured through the number of sentences and mean sentence length, while n-gram diversity and compression ratio provide indicators of repetition and information density. All metrics are computed at the document level and then averaged across reports. Metrics definitions are provided in Appendix D.

5.3 Downstream Evaluation (Information Extraction)

We evaluate the impact of our approach on Information Extraction by comparing models fine-tuned on the top-ranked datasets we created against the same models in both 0-shot and 3-shot settings.

For each training instance, the input consists of a prompt containing a pathology report, the target attributes to be extracted, and, when applicable, their possible values or valid ranges. The expected output is a JSON object containing the extracted values for the corresponding attributes.

In the 3-shot setting, we selected three representative examples covering complementary clinical scenarios: (i) a biopsy case with an in situ tumor, characterized by many missing target values; (ii) a tumorectomy case with lymph-node information, including known numbers of sampled/examined nodes and positive/affected nodes; and (iii) a mastectomy case. This selection was intended to expose the model to different specimen types, varying levels of information completeness, and different lymph-node reporting configurations.

We fine-tune the models using LoRA adaptation and evaluate them on a held-out test set of 377 real breast-cancer pathology reports. These reports were manually annotated by domain experts with gold-standard labels for the 13 target attributes.

Rank	Datasets	Training dataset size	Faithf. Accuracy (%)	High + Perfect docs	Perfect docs
1	$\mathbf{D}_{Mx_Filter}$	561	**96.68**	**998**	674
2	$\mathbf{D}_{MG_RAG_Filter}$	662	96.66	993	**689**
3	$\mathbf{D}_{Mx_RAG_Filter}$	408	96.64	990	*688*
4	$\mathbf{D}_{MG_Filter}$	690	95.97	973	605
5	$\mathbf{D}_{MG}$	1062	94.63	915	485
6	$\mathbf{D}_{MG_RAG}$	1062	94.44	900	490
7	$\mathbf{D}_{Mx}$	1062	93.62	827	489
8	$\mathbf{D}_{Mx_RAG}$	1062	90.57	666	355
T	$\mathbf{D}_{MG_Base}$	1062	96.08	965	629

Table 2: **Ranking of the datasets ($\mathbf{D}^{S1,8}$) produced by the eight MEDGEMMA student models trained on the output of the four teacher models with and without filtering.** The MG_base row indicates the results for the dataset produced by the base MEDGEMMA Model.

In this information extraction setting, the training data consists of the best student-generated outputs, which are used as supervision for LoRA-based fine-tuning. This experimental design allows us to examine whether teacher configurations that perform well for text generation also provide useful supervision for downstream information extraction.

6 Experiments

Attribute	F.Acc MG(%)	F.Acc Mx(%)
diag_test	99.81	99.91
lymph_nodes_taken	94.82	88.70
lymph_nodes_aff	98.49	88.32
extracapsular_spread	94.26	75.99
tumor_size	94.35	95.57
estrogen_receptor	99.44	98.40
progest_receptor	100.00	98.12
lymphov_invasion	90.58	73.73
her2	92.66	87.48
clear_margins	83.71	90.96
ki67	100.00	99.44
morphology	94.44	97.74
grade	96.61	97.83
Macro-average	**95.32**	**91.71**

Table 3: **Deciding on the Student Model.** We compare the faithfulness accuracy of MedGemma-27B (MG) and Mixtral-8×7B (Mx).

6.1 Teacher Models

To increase the diversity of the generated distillation data, we consider four teacher models, comprising two LLMs and two Retrieval-Augmented Generation (RAG) variants of these models. For the LLMs, we use Mixtral-8×7B (MIXTRAL, (Jiang et al., 2024)) for its strong French generation capabilities and MedGemma-27B (MEDGEMMA, https://huggingface.co/google/medgemma-27b-it) for its domain-specific medical expertise. Initial experiments with Qwen3-30B-A3B as a large, instruction-tuned generalist model, were not convincing and were therefore not pursued.

The RAG variants (MX_RAG and MG_RAG) retrieve examples from a database of pre-instantiated medical templates originally used by physicians to accelerate pathology report writing. Rather than retrieving raw templates directly, we first convert them into complete, medically plausible report examples covering the relevant template alternatives. These instantiated reports are then used as retrieval examples during generation (the top-2 most similar instantiated templates). Appendix B describes the structure of the original templates and the procedure used to instantiate them.

All four teacher models were applied to the AV_TEACHER dataset, which contained 1,062 records, each represented as a set of 13 clinical attribute–value pairs. Inference is done using deterministic beam search with four beams; sampling was disabled and temperature was set to 0.

6.2 Student Model

We use a single strong student model to isolate the effects of teacher choice, filtering, and dataset combination. We evaluated MedGemma-27B and Mixtral-8×7B as candidates and selected MedGemma-27B because it achieved higher macro-averaged faithfulness accuracy (Table 3). The MedGemma-to-MedGemma condition is treated as self-distillation.

All experiments use deterministic greedy decoding with sampling disabled, temperature set to 0, and a single beam. This ensures that output variation reflects differences among teachers rather than decoding randomness.

We train two types of student models: S^1, trained on the output of individual teachers, to study teacher choice and filtering; and S^2, trained on

combined teacher outputs, to study dataset combination.

Student Models 1 (S^1). We train the student model (MEDGEMMA) on eight datasets: four datasets generated by the four teacher models (MEDGEMMA, MIXTRAL, MG_RAG, MX_RAG) and the same four datasets filtered for faithfulness accuracy using GPT 5.1 as LLM judge to determine the proportion of input attribute–value pairs correctly expressed in each generated report (cf. Section 5). When filtering, we only retain the data–report pairs with perfect faithfulness accuracy, i.e., where all input attribute–value pairs are correctly expressed in the generated text. We then perform inference on AV_STUDENT, thereby generating eight data–report datasets ($D^{S1,8}$).

This allows us to assess the impact of data distillation (how does the student model compare with its teacher?), the teacher model (which teacher generates the most accurate texts?), and filtering (does filtering improve faithfulness accuracy?).

Student Models 2 (S^2). Intuitively, datasets generated by different models should help improve diversity while using multiple datasets for training might improve learning. To explore this hypothesis, we first rank the datasets in $D^{S1,8}$ by faithfulness accuracy score and progressively increase the training set size, starting from data used to train the best-performing student model and incrementally adding data from the next best models. We consider four combinations (progressively including the five best datasets in D^S_m) and perform inference on AV_STUDENT. This permits assessing the tradeoff between data quantity and data quality.

7 Results and Discussion

Accuracy varies with the teacher model. Table 4 shows the faithfulness accuracy of the datasets created by MEDGEMMA and the four students. The best results are obtained by MEDGEMMA and its self-distilled student model MEDGEMMA^S, highlighting the good ability of MEDGEMMA to process medical text. The RAG models generally underperform their direct generation counterpart suggesting that the retrieved data confuses rather than improves text generation.

A detailed breakdown of the teacher-model results across all attributes is provided in Appendix E.1, including Tables 8 and 7. Overall, these results show that distillation alone is insufficient to produce high-quality synthetic data.

Attribute	MG^T	MG^S	MGR^S	Mx^S	MxR^S
diag_test	100.00	98.77	99.91	100.00	100.00
LNT	94.63	93.03	97.36	87.18	92.74
LNA	98.77	99.81	95.48	91.42	90.95
XCS	94.63	93.12	90.57	87.09	80.58
TS	99.25	99.81	98.96	98.02	73.70
ESR	99.25	100.00	100.00	99.34	98.96
PR	100.00	100.00	100.00	98.49	99.72
LI	90.01	86.33	92.27	67.86	68.05
her2	95.10	83.13	99.34	94.82	98.87
CM	85.11	84.83	82.94	97.27	81.24
ki67	100.00	99.91	100.00	99.81	100.00
morpho	96.32	94.34	95.29	96.98	95.76
grade	96.04	97.08	75.59	98.77	96.89
Average	**96.08**	**94.63**	**94.44**	**93.62**	**90.57**

Table 4: Faithfulness accuracy of the datasets generated by the student models (superscript S) vs the MedGemma teacher model (MG^T). See Table 3 for the expanded names of the attributes.

High Quality Data matters more than Quantity. Table 2 shows the ranking in terms of faithfulness accuracy of the datasets generated by the student models S^1, trained on the eight filtered and unfiltered datasets generated by the four teacher models. Detailed results per attribute are provided in Appendix E.2, and Figure 2 plots training data size against faithfulness accuracy. To assess whether the benefit of filtering extends beyond the faithfulness metric used for data selection, we also evaluate the filtered and unfiltered datasets on downstream information extraction. The results are also reported in Appendix E.2.

Overall, the filtered models achieve higher faithfulness accuracy despite being trained on substantially smaller datasets. The strongest result is obtained by MIXTRAL_FILTER trained on 561 examples, with 96.68% faithfulness accuracy. Similarly, MG_RAG_FILTER and MX_RAG_FILTER reach 96.66% and 96.64%, respectively, while using only 662 and 408 training examples. In contrast, the unfiltered models are all trained on 1,062 examples but achieve lower accuracy, ranging from 90.57% to 94.63%. This suggests that filtering improves data quality enough to offset, and in this case surpass, the reduction in training set size. And more generally, that data quality rather than quantity is key in learning accurate text generation models.

Training on Filtered Data does not reduce Lexical Diversity. The lexical and diversity statistics are shown in Table 5. The main question here is:

Which model outputs look most similar to gold in length, lexical variety, repetition, structure, and fluency? Overall, the results show that filtering does not reduce lexical diversity. In several cases, the filtered variants are as close as, or closer than, their unfiltered counterparts to the gold references. For the MG models, MEDGEMMA_FILTER remains very close to MEDGEMMA across most metrics, with slightly higher length, unique-token count, and MTLD. This suggests that filtering preserves the lexical behaviour of the base model while improving output quality. For the Mx models, the effect of filtering is more pronounced when combined with retrieval: MX_RAG_FILTER is the closest system to gold in terms of length and unique-token count, and it also substantially improves MTLD compared with MX_RAG.

Comparing each RAG model with its non-RAG counterpart shows that retrieval changes the output profile, but not always in the same direction. For MG, retrieval produces shorter outputs: MG_RAG is considerably shorter than MEDGEMMA, with fewer sentences and fewer unique tokens. It also has higher n-gram diversity and higher MTLD. Since TTR is sensitive to text length, its higher value for the shorter MG_RAG outputs is treated as descriptive only and is not used as independent evidence of greater lexical diversity. Overall, the MTLD and n-gram results suggest a change in lexical and surface-form variation, together with a shift away from the gold length distribution. The same pattern appears for MG_RAG_FILTER compared with MEDGEMMA_FILTER.

For Mx, retrieval has a more mixed effect. MX_RAG is shorter than MIXTRAL and has substantially lower MTLD. In contrast, MX_RAG_FILTER improves over MIXTRAL_FILTER on several measures, including length, unique-token count, n-gram diversity, compression ratio, sentence count, and MTLD. Although TTR also increases, we do not rely on it for the main lexical-diversity interpretation.

Taken together, these results support the hypothesis that training on filtered data improves output quality without reducing lexical diversity. Filtering generally preserves or improves lexical richness, and in the best case, MX_RAG_FILTER, it produces outputs that are closer to the gold reports in terms of length, unique-token count, and MTLD. The effect of retrieval is less consistent. For the MG models, retrieval increases MTLD and n-gram diversity while producing

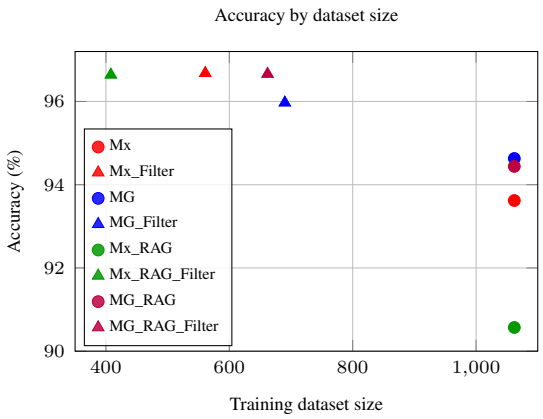


Figure 2: **Impact of Filtering.** Models trained on filtered data (Triangles) achieve better results with less data.

substantially shorter outputs. For the Mx models, its effect depends on whether filtering is applied: MX_RAG has lower MTLD than MIXTRAL, whereas MX_RAG_FILTER improves over MIXTRAL_FILTER on several lexical and structural measures. Overall, retrieval changes the lexical profile of the generated reports, but its effect on diversity is model-dependent and should be interpreted jointly with output length.

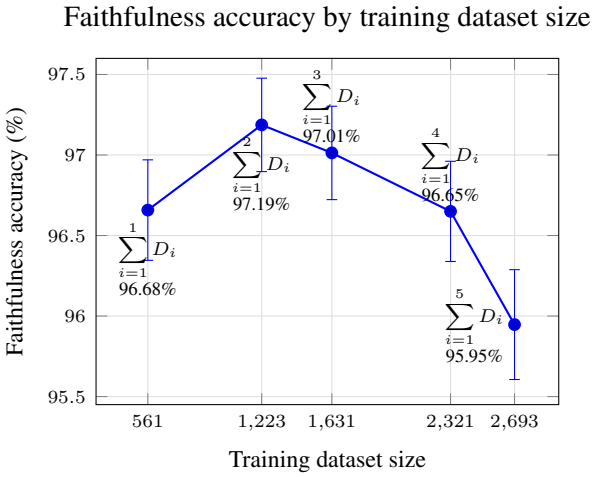


Figure 3: **Combining Datasets: Quantity–Quality Trade-off.** Faithfulness accuracy is shown as a function of cumulative training-set size. Error bars indicate document-level bootstrap 95% confidence intervals based on 10,000 resamples of the held-out reports. Datasets are added according to their faithfulness-accuracy ranking: $D_1 = \mathbf{D}_{Mx_Filter}$, $D_2 = \mathbf{D}_{MG_RAG_Filter}$, $D_3 = \mathbf{D}_{Mx_RAG_Filter}$, $D_4 = \mathbf{D}_{MG_Filter}$, and $D_5 = \mathbf{D}_{MG}$.

Data Quantity vs. Data Quality. Next, we investigate whether combining the training data pro-

model	Gold	MG	MG_Filter	MG_RAG	MG_RAG_Filter	Mx	MxFilter	Mx_RAG	Mx_RAG_Filter
Lexical Volume									
unique_tokens	160.00	137.00	138.00	116.00	118.00	131.00	113.00	120.00	**156.00**
text length	266.00	229.00	**232.00**	160.00	169.00	228.00	214.00	194.00	**232.00**
Lexical Diversity									
mtld	97.17	94.98	**96.65**	135.20	138.92	99.46	79.13	45.81	95.53
ttr	0.58	0.62	0.62	0.72	0.73	**0.58**	0.53	0.59	0.66
Repetition									
ngram_diversity	0.84	0.86	0.87	0.91	0.91	**0.84**	0.79	**0.84**	0.88
compression_ratio	0.54	**0.52**	**0.52**	0.59	0.59	0.50	0.47	**0.56**	0.57
Text Structure									
nb_sentences	36.00	**29.00**	**29.00**	23.00	23.00	23.00	21.00	24.00	28.00
mean_sent_len	7.30	7.93	7.91	**7.43**	7.04	10.11	9.83	8.00	8.10

Table 5: **Comparing the Linguistic Diversity of the Datasets produced by the Student Models.** Gold denotes the real pathology-report corpus used as the reference distribution. Median scores are reported.

duced by different teacher models improves the performance of new student models. Starting from the dataset yielding the best results (D_{Mx_Filter}, Table 2), we incrementally extend the training data by adding the training data of the next best model removing duplicates.

Figure 3 shows that combining the two top-ranked datasets yields the highest faithfulness accuracy, reaching 97.19% (95% bootstrap CI: 96.90–97.47), a significant improvement of 0.53 percentage points over Top1 (95% CI: 0.16–0.90). Adding Top3 or Top4 does not produce a statistically clear change, whereas Top4 performs significantly worse than Top2 and Top5 further decreases performance. These results support selective aggregation of high-quality teacher-generated data: the best student is obtained from a carefully chosen combination rather than from the largest training set.

Downstream Evaluation: Impact on Information Extraction. We evaluate the impact of synthetic data on downstream information extraction by fine-tuning Mistral-7B-Instruct-v0.3 and MEDGEMMA-27B. Detailed attribute-level results are reported in Appendix F, Tables 19 and 18.

Overall, fine-tuning on synthetic data improves information extraction performance for both models. For MEDGEMMA-27B, the gains are modest: the best fine-tuned model improves over the 0-shot baseline from 0.86 to 0.90 in F1 score, showing the strong capacity of this LLM on medical data. In contrast, for the smaller model, synthetic data leads to substantial gains: the best fine-tuned variant improves from 0.57 to 0.86 in F1 compared with the 0-shot baseline, matching the zero-shot performance of MedGemma-27B. We again observe that combining data from different sources (either by merging outputs from the top two or three student models, or by integrating data from the best student models with the teacher model), further improves performance. Overall, the results indicate that MedGemma has substantially stronger medical priors, while DDKD training provides Mistral-7B with sufficient task-specific supervision to close most of the performance gap.

8 Conclusion

In this work, we investigated data-driven knowledge distillation from four teachers, along with data filtering and data combination, to generate synthetic breast cancer patient reports. Our results show that, in the evaluated data regime, filtered higher-quality data outperforms larger unfiltered datasets in terms of faithfulness, highlighting the importance of careful data curation for producing effective synthetic datasets.

Combined with the use of a tabular data synthesiser, our report generation framework enables the creation of paired data-report datasets which we show support the training of effective information extraction models.

More broadly, our findings suggest that synthetic data can be used effectively to train compact information extraction models for the medical domain, which is valuable given that many healthcare organisations lack access to large-scale computational resources.

9 Limitations

Model coverage. We evaluate a limited set of model families. For text generation from tabular data, we primarily rely on MedGemma 27B-Text-IT and Mixtral 8x7B-Instruct as teacher models, with and without retrieval-augmented generation, and MedGemma as the student model. For

Information Extraction, we evaluate models fine-tuned on the generated data, mainly comparing MedGemma 27B-Text-IT and Mistral 7B-Instruct as base models. Although fine-tuned Mistral 7B-Instruct achieves performance comparable to base MedGemma 27B-Text-IT in this setting, further experiments with a broader range of model architectures and sizes would be needed to assess the robustness of these findings.

Student model for report generation We use a single strong report-generation student to isolate the effects of teacher source, filtering, and dataset combination. Consequently, the experiments do not establish how these effects vary with student architecture or capacity. Future work should evaluate smaller and non-domain-specific students and compare the resulting trade-offs among generation quality, computational cost, and downstream utility.

Absence of a real-data oracle. The available structured patient data are not aligned one-to-one with individual pathology reports, as a patient may have multiple reports and this alignment was not retained. Consequently, these data cannot be used directly to train a reliable real-data Information Extraction baseline without introducing potentially incorrect supervision. Future work should construct a larger expert-annotated training set and use it to quantify the remaining gap between synthetic and real-data supervision.

Clinical Scope Finally, this work focuses only on French-language breast cancer cases and on a predefined set of 13 clinical attributes. As a result, the generalisability of the framework to other cancer types, languages, institutions, or broader clinical domains remains to be evaluated. Extending the approach to additional oncology settings, other medical specialties, and more heterogeneous clinical records would be an important direction for future work.

Scale of the quality–quantity analysis Our experiments involve relatively modest training-set sizes. At larger scales, the balance between data quality and quantity may depend on the structure of the errors: diverse and approximately unbiased noise may partly average out, whereas systematic generation errors may be reinforced. Evaluating this trade-off over substantially larger synthetic datasets is an important direction for future work.

10 Ethical considerations

The study received approval from the institution's internal bodies. The mandatory declaration was completed with the French regulatory authority in accordance with Reference Methodology MR-004 (ref. MR-004, no. 28088412).

Acknowledgments

We thank the anonymous reviewers for their feedback. This work received government funding managed by the French National Research Agency under France 2030, reference number "ANR-23-IACL-0004" (AI Chair Gardent: "Semantically Consistent LLM Based Text Generation"). This work was granted access to the HPC resources of IDRIS under the allocation AD011016561 made by GENCI. The authors thank SELARL U2R for providing the data and supporting this research.

11 Use of AI Assistant

We used ChatGPT for help with local rewriting of short English text fragments.

References

Arno Appenzeller, Moritz Leitner, Patrick Philipp, Erik Krempel, and Jürgen Beyerer. 2022. Privacy and utility of private synthetic data for medical data analyses. *Applied Sciences*, 12(23):12320.

Nitay Calderon, Subhabrata Mukherjee, Roi Reichart, and Amir Kantor. 2023. A systematic study of knowledge distillation for natural language generation with pseudo-target training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14632–14659.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Peter Deutsch. 1996. Deflate compressed data format specification version 1.3. RFC 1951.

Georgi Ganev, Meenatchi Sundaram Muthu Selva Annamalai, Sofiane Mahiou, and Emiliano De Cristofaro. 2025. The importance of being discrete: Measuring the impact of discretization in end-to-end differentially private synthetic data. *arXiv preprint arXiv:2504.06923*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188.

Bastian Grothey, Jan Odenkirchen, Adnan Brkic, Birgid Schömig-Markiefka, Alexander Quaas, Reinhard Büttner, and Yuri Tolkach. 2025. Comprehensive testing of large language models for extraction of structured data in pathology. *Communications Medicine*, 5(1):96.

Werner O Hackl, Sabrina B Neururer, Stefan Richter, Hasan Taha, Helmut Muehlboeck, Christoph Hickmann, Patricia Gscheidlinger, Martin Danler, Marco Schweitzer, Marko Ueberegger, and 1 others. 2025. Development of a synthetic oncology pathology dataset for large language model evaluation in medical text classification. In *dHealth 2025*, pages 215–220. IOS Press.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in french. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Zdeněk Kasner and Ondřej Dušek. 2024. Beyond traditional benchmarks: Analyzing behaviors of open llms on data-to-text generation. *arXiv preprint arXiv:2401.10186*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, Annenberg School for Communication, University of Pennsylvania.

Chanseo Lee, Sonu Kumar, Kimon A Vogt, Muhammad Munshi, Panindhra Tallapudi, Antonia Vogt, Hamzeh Awad, and Wasim Khan. 2025. Democratizing cost-effective, agentic artificial intelligence to multilingual medical summarization through knowledge distillation. *Scientific Reports*, 15(1):27619.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Alejandra Lorenzo, Adrien Coulet, and Claire Gardent. 2026. Privacy-Preserving Generation of Synthetic Pathology Reports for Information Extraction. In *AIME 2026 - International Conference on Artificial Intelligence in Medicine*, Ottawa, Canada.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Elisabetta Munzone, Antonio Marra, Federico Comotto, Lorenzo Guercio, Claudia Anna Sangalli, Martina Lo Cascio, Eleonora Pagan, Davide Sangalli, Ilaria Bigoni, Francesca Maria Porta, and 1 others. 2024. Development and validation of a natural language processing algorithm for extracting clinical and pathological features of breast cancer from pathology reports. *JCO Clinical Cancer Informatics*, 8:e2400034.

Anthony Nguyen, Michael Lawley, David Hansen, and Shoni Colquist. 2012. Structured pathology reporting for cancer from free text: Lung cancer case study. *electronic Journal of Health Informatics*, 7(1):e8.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Phillip Park, Yeonho Choi, Nayoung Han, Ye-Lin Park, Juyeon Hwang, Heejung Chae, Chong Woo Yoo, Kui Son Choi, and Hyun-Jin Kim. 2025. Leveraging natural language processing for efficient information extraction from breast cancer pathology reports: Single-institution study. *Plos one*, 20(2):e0318726.

Jake Probert, David Dodwell, John Broggio, Jackie Charman, Mitch Dowsett, Amanda Kerr, Paul McGale, Carolyn Taylor, Sarah C Darby, and Gurdeep S Mannu. 2023. Ki67 and breast cancer mortality in women with invasive breast cancer. *JNCI Cancer Spectrum*, 7(5):pkad054.

Fahad Shahid, Min-Huei Hsu, Yung-Chun Chang, and Wen-Shan Jian. 2025. Using generative ai to extract structured information from free text pathology reports. *Journal of Medical Systems*, 49(1):36.

Andrey Sidorenko, Michael Platzer, Mario Scriminaci, and Paul Tiwald. 2025. Benchmarking synthetic tabular data: A multi-dimensional evaluation framework. *arXiv preprint arXiv:2504.01908*.

Yifei Song, William Soto Martinez, Anna Nikiforovskaya, Evan Parker Kelly Chapple, and Claire Gardent. 2025. Multilingual verbalisation of knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1111–1162. Association for Computational Linguistics.

Hendrik Šuvalov, Mihkel Lepson, Veronika Kukk, Maria Malk, Neeme Ilves, Hele-Andra Kuulmets, and Raivo Kolde. 2025. Using synthetic health care data to leverage large language models for named entity recognition: development and validation study. *Journal of Medical Internet Research*, 27:e66279.

Fiona J. Tweedie and R. Harald Baayen. 1998. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

Elizabeth Geena Woo, Michael C Burkhart, Emily Alsentzer, and Brett K Beaulieu-Jones. 2025. Synthetic data distillation enables the extraction of clinical information at scale. *npj Digital Medicine*, 8(1):267.

Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A Inan, Harsha Nori, Haotian Jiang, Huishuai Zhang, Yin Tat Lee, and 1 others. 2024. Differentially private synthetic data via foundation model apis 2: Text. *arXiv preprint arXiv:2403.01749*.

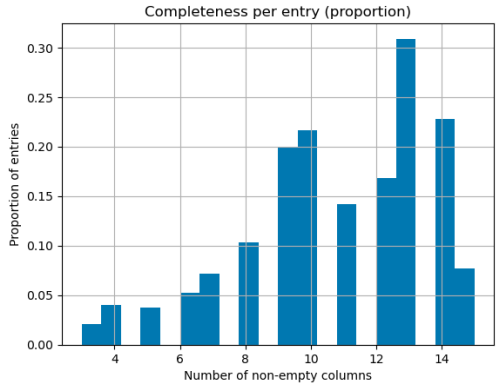
Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional GAN. *Advances in neural information processing systems*, 32.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342.

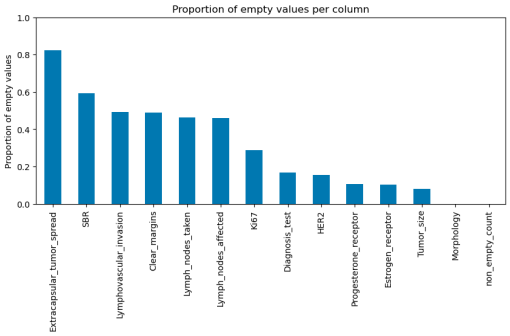
A Synthetic Tabular Data Generation

A.1 Dataset

We used the attributes listed in Table 1.



(a) Non-empty columns per row



(b) Empty values per column

Figure 4: Data completeness: (a) Distribution of non-empty columns per entry. (b) Percentage of empty values per column.

Data completeness is summarized in Figure 4, which shows both the distribution of non-empty columns per entry (Figure 4a) and the proportion of empty values per column (Figure 4b).

A.2 Data preprocessing

Prior to synthetic tabular data generation, a comprehensive data preprocessing phase was performed to ensure data quality, interpretability, and help further with privacy protection.

A.3 Data cleaning and formatting

All categorical attributes (e.g., diagnostic method or sbr grade) were standardized into consistent nomenclature to ensure harmonized encoding. Also, numerical columns were converted to appropriate numeric types and encoded where necessary for downstream model compatibility.

A.4 Manual Discretization of Continuous Attributes

All continuous attributes describing tumor and pathological characteristics (such as tumor size, ki67 , and number of lymphs removed), were discretized into clinically meaningful ranges prior to synthesis.

The discretization was defined manually by domain experts based on established medical thresholds (e.g., tumor size intervals reflecting T1–T4 stages, Ki-67 proliferation index ranges).

Importantly, bin edges were determined independently of the dataset (they were not computed from observed minimums, maximums, or quantiles), thereby preventing any data-dependent information leakage.

Because the discretization process is data-independent, it does not consume any portion of the differential privacy budget (ϵ), and all privacy guarantees apply solely to the synthesis step.

- Test used for diagnosis: This attribute was simplified to reduce granularity and improve both interpretability and privacy. Instead of using all specific procedural descriptions (e.g., “tumorectomie avec curage sentinelle”, “tumorectomie avec curage axillaire”, “exérèse chirurgicale”, etc.), the values were categorized under broader, clinically meaningful groups: tumorectomie, mammectomie and biopsie. This recoding ensures that similar procedures are grouped together, reduces the number of unique categories, and maintains clinical relevance while enhancing privacy in the synthetic dataset.

- Morphology (according to the International Classification of Diseases for Oncology): already discrete.

- Scarff-Bloom-Richardson (SBR) grade: "sbr1", "sbr2","sbr3"

- Tumor size: instead of using the raw tumor size values (measured in millimeters), we categorize them into clinically relevant intervals based on the T classification: 0, 1–5, 6–10, 11–20, 21–50, and 50+ mm.

- Lymph Node attributes: Two attributes were included in the dataset describing lymph node evaluation:

  – Number of lymph nodes removed, and

– Number of lymph nodes affected

Both were discretized according to the standard N classification used in oncology. The discretization reflects clinically meaningful categories that are commonly used in tumor staging and pathology reporting: 0, 1-3, 4-9, +10.

- Clear margins: already discrete, described as oui, non.
- Lymphovascular Invasion: already described as pressent (oui) or absent (non)
- Extracapsular tumor spread: already described as pressent (oui) or absent (non)
- Receptors for the hormone estrogen (RE): already described as 'positifs', 'négatifs'
- Receptors for the hormone progesterone (RP): same as RE.
- The HER2 (CerB2) status attribute already described as 'positifs', 'négatifs'
- The Ki-67 proliferation index was discretized into clinically meaningful ranges to align with established prognostic groupings in breast cancer pathology. Discretization was based on the classification used in Probert et al. (2023) and complemented by expert clinical suggestions. Following this reference, Ki-67 values were categorized into the following intervals, reflecting commonly accepted clinical thresholds for tumor proliferation activity: 0-5, 6-10, 11-20, 21-30, 31-40, 41-100.

A.5 Handling Rare Categories

After discretization, all attributes were treated as categorical. To protect privacy, any category with fewer than five cases was removed from the dataset. This measure is especially important for morphology and other rare clinical features, as it reduces the risk of re-identification while ensuring adequate counts for robust statistical analysis. In total, the final dataset is comprised of 3,880 patient records.

A.6 Validation Constraints

After discretization and recoding, medical validation rules were applied to ensure that all rows in the dataset represent plausible clinical cases. The following constraints were enforced based on expert knowledge:

- SBR grade: Assigned only if the tumor morphology is invasive (infiltrant).
- Vascular emboli: Recorded only if the tumor morphology is invasive (infiltrant).
- Lymph node dissection: 1- Number of lymph nodes removed and affected recorded only if the tumor is invasive (infiltrant).

  2- The number of lymph nodes removed must be greater than or equal to the number of affected lymph nodes.
- Extracapsular tumor spread: Recorded only if at least one lymph node is affected.
- Biopsies should not have tumor size or margins value assigned;

Rows violating any of these constraints were excluded from the final dataset. This step ensures that synthetic data generation is based on medically plausible and internally consistent cases, enhancing both the utility and credibility of the resulting synthetic dataset.

A.7 Choice of privacy budget and Evaluation of Similarity to the real data and of Privacy Preservation

In differential privacy, a lower ϵ enforces stronger privacy guarantees at the cost of reduced data utility, whereas a higher ϵ improves accuracy but weakens privacy protection.

Following Appenzeller et al. (2022), we evaluated multiple values of ϵ to identify an appropriate trade-off between privacy and utility. For each value of ϵ, we conducted three independent runs and report the average results in Table 6 using Accuracy, DCR (Distance to Closest Record) (Sidorenko et al., 2025) and number of identical matches. *Accuracy (Acc.)* (Sidorenko et al., 2025) measures the agreement between real and synthetic data, measured across univariate (single column), bivariate (column pairs), and trivariate (column triplets) distributions via averaged L1 distance (higher is better). The *DCR* score shows the proportion of synthetic records that are closer to some training record than to any holdout record (lower is better). Accuracy measures similarity to the real data while DCR and *Number of identical matches (Id.Mtchs)* measure privacy preservation.

Among the evaluated privacy budgets, $\epsilon = 3$ provided the most favorable observed privacy–utility trade-off in our experimental setting. Across

ϵ	Acc. (%, ↑)	DCR (%, ↓)	Id.Mtchs ↓
2.0	67.57 ± 0.49	58.47 ± 3.82	0.03 ± 0.06
3.0	79.60 ± 0.53	52.53 ± 0.57	0.47 ± 0.15
4.0	79.77 ± 0.93	52.40 ± 1.30	2.60 ± 0.10
5.0	72.37 ± 0.80	52.23 ± 2.61	8.47 ± 0.67

Table 6: Experimental results for PATE_CTGAN across different ϵ values (averaged over 3 runs). Values are reported as mean ± standard deviation.

three independent runs, accuracy increased from $67.57 \pm 0.49\%$ at $\epsilon = 2$ to $79.60 \pm 0.53\%$ at $\epsilon = 3$, while the proportion of identical matches remained low at $0.47 \pm 0.15\%$. Increasing the budget to $\epsilon = 4$ produced a similar accuracy of $79.77 \pm 0.93\%$, but increased identical matches to $2.60 \pm 0.10\%$. At $\epsilon = 5$, identical matches increased further to $8.47 \pm 0.67\%$, while accuracy decreased to $72.37 \pm 0.80\%$. DCR values were broadly similar for $\epsilon = 3$, $\epsilon = 4$, and $\epsilon = 5$, and therefore did not provide additional evidence in favor of a larger privacy budget.

This pattern is consistent with prior work reporting a similar privacy–utility trade-off for differentially private synthetic medical data (Appenzeller et al., 2022). We therefore selected $\epsilon = 3$ for the subsequent experiments. This choice is specific to the evaluated dataset, synthesis mechanism, and empirical diagnostics, and should not be interpreted as a universal cutoff or as evidence that $\epsilon = 3$ is intrinsically a strong privacy guarantee. Identical-match rates and DCR are empirical privacy diagnostics that complement, but do not replace, the formal differential-privacy guarantee.

B Templates

For the RAG inference we rely on a collection of 328 report templates used by clinicians for the generation of breast pathology reports. Figure 5 shows an example with the templates organisation. The first part contains metadata about the template, followed by the corresponding text.

The templates were not fully standardized. The fields to be modified were not always marked in the same way: in some cases they were indicated with #, while in others they were not explicitly marked. Moreover, the meaning of # was inconsistent. Sometimes it marked a phrase to be instantiated, sometimes a phrase that could be removed, and sometimes the template simply contained several optional phrases listed one after another.

Code : **GSBK03** - Abrégé : P_ adénocarcinome canalaire infiltrant de grade II (ER) - Type : X - **ADICAP :** BHGSA7B2

Examen microscopique
Il montre un adénocarcinome représenté par une prolifération épithéliale d'organisation glandulaire et trabéculaire. La différenciation glandulaire représente environ 50 % de la tumeur (score 2). Le pléomorphisme nucléaire est modéré (score 2). L'index mitotique est de moins de 10 mitoses pour 10 champs x Bx40 Olympus (score 2).
Conclusion
Adénocarcinome mammaire de type canalaire infiltrant moyennement différencié.
Grade II de Scarff et Bloom (score 6).
Etude immuno-histochimique (procédure Dako Omnis Flex Module Omnis) :
· Récepteurs hormonaux :
Œstrogène (GA084/Dako Clone EP1 Mono) :
Réaction négative.
Réaction positive : # % de cellules exprimées. Intensité : #
Progestérone (GA090/Dako Clone PgR 1294 Mono) :
Réaction négative.
Réaction positive : # % de cellules exprimées. Intensité : #
· Statut Her2Neu (A048529-2/ Dako Poly) :
Score 0 : absence de marquage ou marquage membranaire incomplet d'intensité faible, = 10% des cellules tumorales.
Score 1+ : marquage membranaire incomplet d'intensité faible, de plus de 10 % des cellules tumorales.
Score 2+ : marquage membranaire incomplet et/ou complet d'intensité faible à modérée de plus de 10 % des cellules tumorales et marquage membranaire complet fort de moins de 10 % des cellules tumorales : FISH en cours.
Score 3+ : marquage membranaire complet fort de plus de 10 % des cellules tumorales.
· KI67 (GA626/Dako Clone MIB-1 Mono) : # %.

Figure 5: Example of original templates organisation: The first part contains metadata about the templates, followed by the corresponding text.

In an initial test, we used the raw templates directly. However, some model outputs still contained unresolved template markers such as #, or alternative labels such as "Positive. Negative" as part of the generated text.

To address this, we pre-processed the templates with GPT-5.4. The goal was to generate as many medically plausible instantiated versions of each template as possible, covering the relevant alternatives present in the original template. The prompt used for this pre-processing step was the following:

```
Tu es un médecin anatomopathologiste expert en
pathologie mammaire.

Lis l'abrégé et le template brut ci-dessous et
produis des exemples textuels instanciés.

OBJECTIF
Retourner EXHAUSTIVEMENT un exemple complet
pour chaque combinaison médicalement possible
des alternatives pertinentes présentes dans
le template.

UTILISATION DE L'ABRÉGÉ
L'abrégé fournit un contexte diagnostique
synthétique important pour interpréter
correctement le template.
Utilise l'abrégé pour guider le type de lésion,
```

```
le type de prélèvement, le caractère bénin ou
malin, et pour construire un Titre et une
Macroscopie cohérents.
En cas d'ambiguïté du template, privilégier
une interprétation cohérente avec l'abrégé.


STRUCTURE OBLIGATOIRE
Chaque compte rendu final doit obligatoirement
contenir :
- une section Titre
- une section Examen macroscopique ou
Macroscopie

Si le template ne contient pas explicitement
l'une de ces sections, tu dois l'ajouter avec
un contenu bref, concret, neutre, cohérent et
médicalement plausible, en restant au plus
proche du template.
- Titre : intitulé médical simple et cohérent,
avec si possible latéralité, type de prélè
    vement
et localisation.
  Exemples: "SEIN GAUCHE : EXÉRÈSE D'UN NODULE
 DU QUADRANT SUPÉRO-EXTERNE", "SEIN DROIT :
 TUMORECTOMIE DU QSE AVEC GANGLION SENTINELLE",

 "MICROBIOPSIE MAMMAIRE 14G DU QSI DU SEIN
 GAUCHE"
- Macroscopie : description brève, plausible
et cohérente avec le prélèvement, la lésion et
la conclusion. Elle peut comporter les
dimensions, poids et aspect de la pièce,
l'aspect de la lésion à la coupe, la distance
à la berge la plus proche, les recoupes
éventuelles et les ganglions éventuels.
Exemple: "Pièce d'exérèse mammaire mesurant
3,5 x 2,8 x 2 cm et pesant 9 g.  À la coupe,
on individualise un nodule bien limité,
ovalaire, blanchâtre, ferme, d'aspect
fasciculé, mesurant 1,6 cm dans son plus grand
axe. La lésion est située à 4 mm de la berge
la plus proche.
Le reste du parenchyme mammaire d'aspect
macroscopique banal."
- Titre et Macroscopie: Ne décrire que les
éléments cohérents avec le template.


ALTERNATIVES PERTINENTES À COMBINER
Alternatives pertinentes à prendre en compte
seulement si elles sont réellement proposées
comme options dans le template :
- type de prélèvement
- présence/absence de ganglions prélevés
- présence/absence de rupture capsulaire
- présence/absence d'emboles vasculaires
- statut RO/RP
- statut HER2
- marges saines ou atteintes
- Ki67 si c'est une vraie alternative
  textuelle
- grade SBR si c'est une vraie alternative
  textuelle

RÈGLES
- Identifier les alternatives même si elles
    sont exprimées de façon non standard ou
    implicite.
- Générer le produit combinatoire complet des
    alternatives pertinentes.
- Exclure toute combinaison médicalement
    impossible.
- Si aucune alternative pertinente n'existe, g
    énérer un seul exemple.
- Ne jamais inventer l'opposé d'une option
    absente du template.
- Une option isolée n'est pas une alternative
    binaire.
- Dans chaque exemple, garder une seule option
     par bloc alternatif et supprimer les
    autres.
- Les variations purement numériques ne créent
     pas de branche, mais doivent être
    instanciées par des valeurs plausibles et
    cohérentes.
- Le texte final doit rester aussi proche que
    possible du template.
- Ne pas inventer de données pronostiques
    majeures absentes du template.

BLOCS EXCLUSIFS
- Les conclusions RO/RP forment un bloc
    alternatif exclusif.
- Les conclusions HER2 forment un bloc
    alternatif exclusif.
- Dans chaque exemple, tout le texte détaillé
    doit être cohérent avec l'option RO/RP et
    HER2 retenue.

CONTRAINTES MÉDICALES
- Le grade SBR s'applique seulement aux
    carcinomes infiltrants.
- Ne pas inventer d'emboles vasculaires pour
    une lésion strictement in situ.
- ganglions_atteints <= ganglions_preleves
- si ganglions_atteints = 0,
    rupture_capsulaire = non
- toute instanciation doit être médicalement
    plausible et cohérente

AJOUTS EN CAS D'ABSENCE DANS LE TEMPLATE
- Si la section Titre est absente, l'ajouter
    avec un intitulé simple, neutre et cohé
    rent avec le type de prélèvement.
- Si la section Macroscopie est absente, l'
    ajouter avec une formulation minimale
    descriptive, neutre, plausible et cohé
    rente avec le prélèvement.
- Ne pas ajouter d'autres sections
    obligatoires non demandées.

SORTIE
Réponds uniquement avec le JSON conforme au
    schéma attendu.
Ne retourne jamais une liste vide.

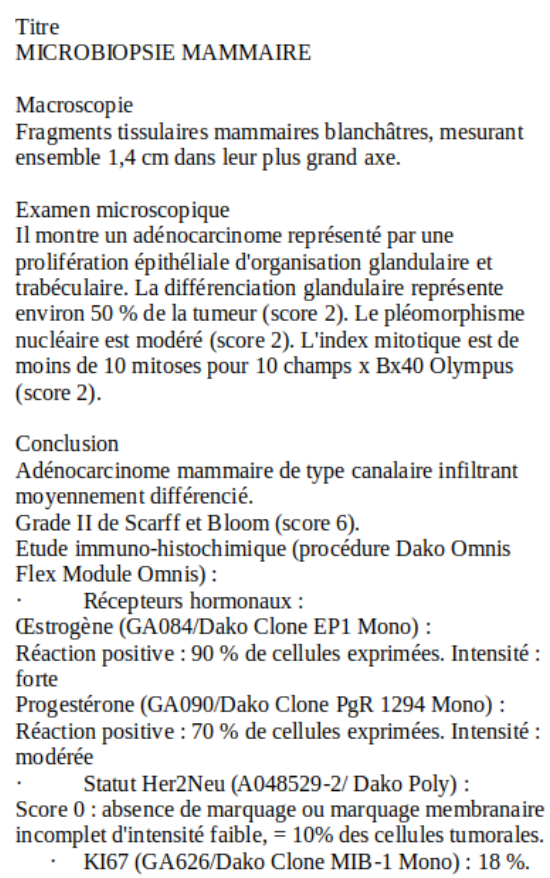
ABRÉGÉ
{ABREGE}
TEMPLATE BRUT
{TEXT}
```

As shown in the prompt, we restricted the set of alternatives to the variables covered in our study in order to avoid an exponential explosion in the number of generated cases.

As a result, we obtained a total of 1480 instanti-

ated templates, which were then used in the RAG-based version of our models.

For indexing, each document was represented by the full instantiated template, enriched with the template name and the anatomical site extracted from the ADICAP code, distinguishing between breast tissue (*sein*) and lymph nodes (*ganglion*). At inference time, the structured input attributes were converted into a short textual query describing the relevant clinical information. The retrieved templates were then provided as contextual examples to guide report generation.



Titre
MICROBIOPSIE MAMMAIRE

Macroscopie
Fragments tissulaires mammaires blanchâtres, mesurant ensemble 1,4 cm dans leur plus grand axe.

Examen microscopique
Il montre un adénocarcinome représenté par une prolifération épithéliale d'organisation glandulaire et trabéculaire. La différenciation glandulaire représente environ 50 % de la tumeur (score 2). Le pléomorphisme nucléaire est modéré (score 2). L'index mitotique est de moins de 10 mitoses pour 10 champs x Bx40 Olympus (score 2).

Conclusion
Adénocarcinome mammaire de type canalaire infiltrant moyennement différencié.
Grade II de Scarff et Bloom (score 6).
Etude immuno-histochimique (procédure Dako Omnis Flex Module Omnis) :
· Récepteurs hormonaux :
Œstrogène (GA084/Dako Clone EP1 Mono) :
Réaction positive : 90 % de cellules exprimées. Intensité : forte
Progestérone (GA090/Dako Clone PgR 1294 Mono) :
Réaction positive : 70 % de cellules exprimées. Intensité : modérée
· Statut Her2Neu (A048529-2/ Dako Poly) :
Score 0 : absence de marquage ou marquage membranaire incomplet d'intensité faible, = 10% des cellules tumorales.
· KI67 (GA626/Dako Clone MIB-1 Mono) : 18 %.

Figure 6: Original template.

C Prompts

C.1 Generation prompt

C.1.1 System Prompt

```
  Tu es un médecin anatomopathologiste
hospitalier expérimenté.
  Tu rédiges des comptes rendus
anatomopathologiques complets, cohérents et
conformes aux standards hospitaliers français.
  Tu corriges automatiquement toute
incohérence médicale, stylistique ou logique.
```

C.1.2 User Prompt Template without RAG

```
TÂCHE:
Rédige un **compte rendu anatomopathologique
complet en français médical** à partir des
données cliniques délimitées
### DONNÉES CLINIQUES ### ci-dessous.

### DONNÉES CLINIQUES ###
{donnees_cliniques}
### FIN DONNÉES CLINIQUES ###


 Tu dois produire exactement les sections
suivantes :
- Titre (indiquer une latéralité et un
  quadrant choisis de façon plausible)
- Examen macroscopique
- Examen microscopique
- Étude immunohistochimique
- Conclusion

----------------------------------------
RÈGLES OBLIGATOIRES
----------------------------------------
1) Cohérence générale
- Toute valeur fixe fournie doit être reprise
  exactement, sans modification.
- Ne jamais faire varier une même donnée entre
  les sections.

2) Données non évaluées
- Toute donnée indiquée comme "non évaluée",
  "non réalisée" ou "non disponible" dans les
  DONNÉES CLINIQUES doit être reprise
  strictement telle quelle.
- Ne jamais la compléter, estimer ou
  interpréter.
- Ne jamais la remplacer par une valeur
  normale ou pathologique (ex: "marges saines",
      "absence d'emboles", etc.).

3) Compléments descriptifs autorisés
- Tu peux enrichir la description
  macroscopique et microscopique avec des
  éléments plausibles, purement descriptifs
  et rédactionnels.
- Tu ne dois jamais modifier, compléter,
  interpréter ni contredire les données
  cliniques fournies.
- Les variables cliniques (taille tumorale,
  ganglions, marges, emboles, récepteurs,
  etc.) doivent être reprises strictement à
  l'identique.

4) Taille tumorale et pièce opératoire
- La taille tumorale ne doit jamais être
  inventée ou estimée si elle n'est pas
  fournie.
- Si elle est fournie, elle doit être
  identique dans la macroscopie, la
  microscopie et la conclusion.
- La taille tumorale doit toujours être
  inférieure ou égale à la taille de la
  pièce opératoire.
- Si poids et dimensions de la pièce sont
  ajoutés, ils doivent être plausibles et
  cohérents entre eux.
```

```
5) Hormonorécepteurs et Ki67
- Si RE ou RP sont positifs, utiliser le
  pourcentage fourni dans les données
  cliniques.
- Si RE ou RP sont négatifs, les rapporter
  comme négatifs sans inventer d'intensité
  absente.
- Une seule valeur par marqueur, sans
  contradiction.

6) Ganglions lymphatiques
- Le nombre de ganglions atteints doit
  toujours être inférieur ou égal au
  nombre de ganglions examinés.
- Le nombre décrit en macroscopie,
  microscopie et conclusion doit être
  identique.
- Si des ganglions sont examinés et qu'il n'y
  a aucune métastase ganglionnaire, indiquer
  0/N dans la conclusion (N = nombre de
  ganglions examinés).
- Si le statut ganglionnaire est non évalué,
  ne pas mentionner de nombre de ganglions ni
  de statut (pas de 0/0, pN0, etc.).
- Ne pas mentionner de rupture capsulaire si
  les ganglions sont négatifs.


7) Grade
- Le grade SBR ne s'applique qu'aux lésions
  infiltrantes.
- Si un score chiffré est présent (ex. 3 + 2
  + 1), calculer le total et en déduire le
  grade :
  - 3 à 5 : grade I
  - 6 à 7 : grade II
  - 8 à 9 : grade III
- Corriger toute incohérence entre score et
    grade.


RÈGLES DE STYLE

- Français médical clair, structuré, homogène,
  sans redondance.
- Fournir uniquement le rapport final.
- N'ajouter aucune explication, justification
  ou note hors compte rendu.
```

Where {donnees_cliniques} is instantiated with the synthetic tabular data described in Appendix A.

C.1.3 User Prompt Template with RAG

```
TÂCHE:
Rédige un **compte rendu anatomopathologique
complet en français médical** à partir des
données cliniques délimitées ### DONNÉES
CLINIQUES ###  ci-dessous.

Les exemples fournis servent uniquement de
modèle de style, de ton et de structure.
Ils doivent être respectés sur la forme,
tout en étant strictement adaptés aux données
cliniques fournies.
N'utilise que les données cliniques fournies
ci-dessous. Tu peux seulement ajouter un
détail rédactionnel générique ou plausible
lorsqu'il est strictement nécessaire pour la
fluidité du compte rendu, à condition qu'il
ne modifie, ne complète abusivement, ne
contredise ni n'interprète les données
cliniques fournies.
Toute donnée inconnue, non évaluée, absente
ou non documentée doit rester inconnue, non
évaluée, absente ou non documentée dans le
texte final. Ne la transforme jamais en donnée
négative, positive ou normale.
Exemple : si les ganglions sont indiqués comme
non évalués, supprime toute mention de
ganglions examinés, de pN0 ou de 0/0.


### DONNÉES CLINIQUES ###
{donnees_cliniques}
### FIN DONNÉES CLINIQUES ###


Tu dois produire exactement les sections
    suivantes :
- Titre  (indiquer une latéralité et un
  quadrant choisis de façon plausible)
- Examen macroscopique
- Examen microscopique
- Étude immunohistochimique
- Conclusion


RÈGLES OBLIGATOIRES

1) Cohérence générale
- Toute valeur fixe fournie doit être reprise
  exactement, sans modification.
- Ne jamais faire varier une même donnée entre
  les sections.

2) Données non évaluées
- Toute donnée indiquée comme "non évaluée",
  "non réalisée" ou "non disponible" dans les
  DONNÉES CLINIQUES doit être reprise
  strictement telle quelle.
- Ne jamais la compléter, estimer ou
  interpréter.
- Ne jamais la remplacer par une valeur
  normale ou pathologique (ex :"marges saines",

  "absence d'emboles", etc.).
- Cette règle s'applique même si les exemples
  contiennent une formulation correspondante.

3) Compléments descriptifs autorisés
- Tu peux enrichir la description
  macroscopique et microscopique avec des
  éléments plausibles, purement descriptifs
  et rédactionnels.
- Tu ne dois jamais modifier, compléter,
  interpréter ni contredire les données
  cliniques fournies.
- Les variables cliniques (taille tumorale,
  ganglions, marges, emboles, récepteurs,
  etc.) doivent être reprises strictement à
  l'identique.

4) Taille tumorale et pièce opératoire
- La taille tumorale ne doit jamais être
  inventée ou estimée si elle n'est pas
  fournie.
```

```
- Si elle est fournie, elle doit être
  identique dans la macroscopie, la
  microscopie et la conclusion.
- La taille tumorale doit toujours être
  inférieure ou égale à la taille de la
  pièce opératoire.
- Si poids et dimensions de la pièce sont
  ajoutés, ils doivent être plausibles et
  cohérents entre eux.

5) Hormonorécepteurs et Ki67
- Si RE ou RP sont positifs, utiliser le
  pourcentage fourni dans les données
  cliniques.
- Si RE ou RP sont négatifs, les rapporter
  comme négatifs sans inventer d'intensité
  absente.
- Une seule valeur par marqueur, sans
  contradiction.


6) Ganglions lymphatiques
- Le nombre de ganglions atteints doit
  toujours être inférieur ou égal au nombre
  de ganglions examinés.
- Le nombre décrit en macroscopie, microscopie
  et conclusion doit être identique.
- Si des ganglions sont examinés et qu'il
  n'y a aucune métastase ganglionnaire,
  indiquer 0/N dans la conclusion (N =
  nombre de ganglions examinés).
- Si le statut ganglionnaire est non évalué,
  ne pas mentionner de nombre de ganglions ni
  de statut (pas de 0/0, pN0, etc.).
- Ne pas mentionner de rupture capsulaire si
  les ganglions sont négatifs.

7) Grade
- Le grade SBR ne s'applique qu'aux lésions
  infiltrantes.
- Si un score chiffré est présent (ex. 3 + 2 +
  1), calculer le total et en déduire le grade:

  - 3 à 5 : grade I
  - 6 à 7 : grade II
  - 8 à 9 : grade III
- Corriger toute incohérence entre score et
  grade.

8) Exemples
- Supprimer tout doublon ou variante
    contradictoire issus des exemples.
- Ne retenir que ce qui est compatible avec
    les données du patient.
- Les exemples servent uniquement à reproduire
     le style rédactionnel, jamais les données
    .
- Si une variable est définie comme absente ou
     non évaluée dans les DONNÉES CLINIQUES,
    ne jamais la réintroduire à partir des
    exemples.


RÈGLES DE STYLE


- Français médical clair, structuré, homogène,
     sans redondance.
- Fournir uniquement le rapport final.
- N'ajouter aucune explication, justification
    ou note hors compte rendu.

### EXEMPLES ###
{RAG_ex}
### FIN EXEMPLES ###
```

Where {donnees_cliniques} is instantiated with the synthetic tabular data described in Appendix A, and {RAG_ex} consist of the Templates obtained by passing this same information through the RAG pipeline described in Appendix B.

C.2 Evaluation Prompts

C.2.1 Faithfulness to the tabular input

In order to evaluate this criterion, we identify first the type of the attribute and whether it has a value assigned in the tabular data.

- Prompt when the value is not provided

```
You are a helpful assistant.

You will receive:

- an ATTRIBUTE, after the token ##ATTRIBUTE##
- a French TEXT, after the token ##TEXT##

Your task is to determine whether the TEXT
    gives a **clear, explicit and evaluable
    result** for this ATTRIBUTE
(e.g. positive, negative, grade, percentage,
    yes, no), and if so, extract the exact
    phrase from the TEXT.

Important:
- The "reference" MUST be copied exactly from
    the TEXT:
  - It must be a contiguous substring of the
    TEXT.
  - You must not add, remove or change any
    words.
  - If you cannot find an exact substring in
    the TEXT, you must answer:
    {{"correct": "no", "reference": "-"}}.

{dict_examples_prompt[att]}

Decision rules:

1. If the TEXT gives a clear, explicit RESULT
    for the ATTRIBUTE
   (e.g. "HER2 3+", "HER2 négatif", "HER2
    surexprimé" for "statut HER2"),
   answer:
   {{"correct": "yes", "reference": "exact
    phrase from the text giving the result"}}

2. If the ATTRIBUTE is mentioned ONLY to say
    it was NOT evaluated or is UNKNOWN
   (e.g. "n'a pas été évaluée", "non testé", "
    non recherché", "non réalisé", "non
    contributif",
       "résultat non interprétable", "inconnu
    ", "pas de résultat disponible"),
   then the result is NOT known. Answer:
```

```
   {{"correct": "no", "reference": "-"}}

3. If the ATTRIBUTE is not mentioned at all in
     the TEXT, and no result can be deduced
    explicitly, answer:
   {{"correct": "no", "reference": "-"}}

4. Never infer or guess a value for the
    ATTRIBUTE from context or other
    information.
   If the explicit result is not written in
    the TEXT, answer:
   {{"correct": "no", "reference": "-"}}

Additional rules:
- Respond ONLY with valid JSON.
- Do NOT add comments or explanations.

Output format:
Return exactly one JSON object, for example:
{{"correct": "yes", "reference": "HER2 négatif
     (score 0)"}}

##ATTRIBUTE##: {attribute_name}
##TEXT##: {texte}
```

- Prompt when the value is provided

```
You are helpful assistant.

You will receive:

- an ATTRIBUTE, delimited by ##ATTRIBUTE##
- a REQUESTED VALUE delimited by ##REQUESTED
    VALUE##
- a French TEXT, delimited by ##TEXT##

Your task is to determine whether the TEXT
    supports the statement:
“ATTRIBUTE = REQUESTED VALUE”.

Follow these steps internally, but output ONLY
     the final JSON:

-------------------------------------------

STEP 1  Determine whether the TEXT contains
    any expression corresponding to the
    ATTRIBUTE (even if wording differs).
    If the ATTRIBUTE is not mentioned at all:
    Return ONLY this JSON:

    {{
      "correct": "no",
      "reference": "-"
    }}
-------------------------------------------

STEP 2  If the ATTRIBUTE is mentioned:
Extract the actual value expressed in the TEXT.


-------------------------------------------

STEP 3  Compare the extracted value to the
    REQUESTED VALUE.
You MUST follow this rule:

- If extracted value == REQUESTED VALUE:
    Return:
    {
      "correct": "yes",
      "reference": "exact phrase from the text
     indicating presence or absence"
    }

- If extracted value != REQUESTED VALUE:
    Return:
    {
      "correct": "no",
      "reference": "-"
    }
IMPORTANT:
When extracted value != REQUESTED VALUE, you
    MUST set "reference": "-"
Do NOT copy any phrase from the text in this
    case.

-------------------------------------------

{dict_examples_prompt[att]}

-------------------------------------------
Additional rules:
- Respond ONLY with a valid JSON.
- Do NOT add comments or explanations.

EXACT output format:
{{
  "correct": "yes" | "no",
  "reference": "exact phrase from the text" |
    "-"
}}


##ATTRIBUTE##: {attribute_name}
##REQUESTED VALUE##: {format_text_att(att,val)}

##TEXT##: {texte}
```

- {dict_examples_prompt[att]}: "For example, in the text: ’MAMMECTOMIE DROITE AVEC CURAGE AXILLAIRE MONOBLOC’, there is a mention to the ###ATTRIBUTE###, and we can extract that the value is ’MAMMECTOMIE’."
- {attribute_name}: The attribute that is being evaluated, for example ’Morphology’.
- {texte}: The synthetic pathological report
- {value}: The value associated with the attribute
- {extra_rules}: Depending on the type of the attribute,
  - If the attribute is boolean, the prompt uses instead the following rules:

```
STEP 2  If the ATTRIBUTE is mentioned:
    Extract the actual value expressed in the
    TEXT:
```

```
    - If the text explicitly expresses
    presence (e.g., “présence de”, “embol(e)
    vasculaire identifié”), extracted value =
    "oui".
    - If the text expresses absence (e.g., “
    absence de”, “aucun”, “pas de”, “sans”),
    extracted value = "non".

-------------------------------------------

STEP 3  Compare the extracted value to the
    REQUESTED VALUE.

You MUST follow this rule:

- If extracted value == REQUESTED VALUE:
    Return:
    {
      "correct": "yes",
      "reference": "exact phrase from the text
     indicating presence or absence"
    }

- If extracted value != REQUESTED VALUE:
    Return:
    {
      "correct": "no",
      "reference": "-"
    }

IMPORTANT:
When extracted value != REQUESTED VALUE, you
    MUST set "reference": "-"
Do NOT copy any phrase from the text in this
    case.
```

C.3 Cost of LLM-as-a-Judge Evaluation

The LLM-as-a-judge evaluation is cost-intensive, as each prompt includes long pathology reports and the evaluation is repeated across multiple experimental settings. Across all test instances, the total evaluation cost amounts to approximately $350.

Training and computational cost. All fine-tuning experiments were performed on a single NVIDIA RTX PRO 6000 Blackwell Server Edition GPU with 96 GB of GPU memory, corresponding to approximately 95 GB of usable VRAM as reported by the system.

Text-generation inference was run on two NVIDIA A100 80 GB GPUs. Depending on the model and prompting configuration, generating the synthetic reports with MedGemma and Mixtral required approximately 7–12 hours. Each LoRA fine-tuning run of MedGemma for text generation took average 6 hours on the RTX PRO 6000 GPU, with peak memory usage up to 88 GB.

For the information-extraction experiments, LoRA fine-tuning of MedGemma-27B-Text required approximately 30–60 minutes per run and used about 60 GB of GPU memory. LoRA fine-tuning of Mistral-7B-Instruct required approximately 30 minutes per run; with peak GPU memory usage of about 18 GB. Inference for the information-extraction models was also performed on the RTX PRO 6000 GPU.

C.4 Human Annotation

The annotation instructions were:
For each attribute, decide only whether the displayed value is correct with respect to the generated report text.

Mark “oui” when the displayed attribute is correct:

- If the displayed value is specified: Mark “oui” only if the report explicitly supports the statement "attribute = displayed value or displayed range".

  The value may be a category, a yes/no value, a number, or a range.

  Example: If the tool shows Ki-67 : 20–30%, mark “oui” if the report explicitly says a Ki-67 value within that range, such as 25%.

- If the displayed value is “non spécifié” (unknown): Mark “oui” only if the report does not contain an explicit evaluable value for that attribute.

  This means the attribute is correct as unknown when the attribute is absent from the report, or when the report says it was not evaluated / not tested / not available / unknown / non interpretable, and no actual value is given.

  Example: If the tool shows "Ki-67 : non spécifié", mark “oui” if there is no Ki-67 value or range in the report.

Mark “non” when the displayed value is not supported by the report:

- the report gives a value but is not correct (value outside the range for ranges, different category for categorical variables, etc.);
- the report contains an explicit value but the tool shows non spécifié;
- the displayed value is inferred or guessed but not explicitly supported by the report.

When marking “oui”, copy in the reference field the exact phrase from the report that supports the value. For non spécifié (unknown), leave the reference empty.

D Diversity Metrics

We compute all descriptive metrics at the document level and then summarize them across reports. The implementation computes the following metrics.

Length. Length is defined as the total number of word tokens in a document. It captures the overall lexical volume of a report.

Unique tokens. Unique tokens correspond to the number of distinct word types in a document. This metric provides a simple measure of vocabulary size, although it is sensitive to document length.

Type-token ratio (TTR). TTR measures lexical diversity as the ratio between the number of unique word types and the total number of word tokens:

$$TTR = \frac{V}{N},$$

where V is the number of unique types and N is the total number of tokens. Higher TTR values indicate greater lexical variety, although TTR is known to be sensitive to text length (Tweedie and Baayen, 1998; McCarthy and Jarvis, 2010).

Measure of Textual Lexical Diversity (MTLD). MTLD is a lexical diversity metric designed to be less sensitive to text length than raw TTR. It computes the average span required for the running TTR to fall below a predefined threshold. Higher MTLD values indicate greater lexical diversity (McCarthy and Jarvis, 2010).

N-gram diversity. N-gram diversity measures the proportion of distinct n-grams in a document. This metric follows the same principle as the Distinct-n measure proposed for evaluating diversity in generated text (Li et al., 2016). In our implementation, we compute distinct n-gram ratios for $n = 1, \ldots, 4$ and report their mean:

$$\text{N-gram diversity} = \frac{1}{4} \sum_{n=1}^{4} \frac{|\text{unique n-grams}|}{|\text{all n-grams}|}.$$

Higher values indicate greater surface diversity and less repeated local phrasing.

Compression ratio. Compression ratio is computed as the ratio between the compressed UTF-8 byte length of a document and its original UTF-8 byte length: We use zlib compression. Lower values indicate that the text is more compressible, which often reflects greater regularity or repetition; higher values indicate less compressible text (Deutsch, 1996).

Number of sentences. We approximate the number of sentences by splitting each document on sentence final punctuation and newlines. This metric captures coarse structural organization, though it may be sensitive to abbreviations, formatting, and nonstandard punctuation.

Mean sentence length. Mean sentence length is defined as the average number of word tokens per sentence. It provides a document-level measure of sentence level structural density.

E DDKD results

E.1 Additional teacher-model results

Tables 7 and 8 provide the full attribute-level faithfulness accuracy of the teacher models on AV_STUDENT and AV_TEACHER, respectively. These results complement the analysis in Section 7 by showing that MedGemma-based models generally achieve higher accuracy than Mixtral-based models.

The impact of RAG is mixed. While MG_RAG slightly improves over MEDGEMMA on AV_STUDENT, it underperforms on AV_TEACHER. In contrast, MX_RAG consistently performs worse than MIXTRAL. Together with the student-model results in Table 4, these findings indicate that neither retrieval nor distillation provides consistent improvements across all models.

E.2 Impact of filtering

Tables 10 to 17 show the detailed faithfulness accuracy results per attribute for the four student models comparing the results when trained on filtered vs unfiltered data.

A notable outlier concerns the *grade* attribute for MG_RAG. In 23.5% of the evaluated cases, MG_RAG generated an incorrect grade while MG generated the correct value. Among these cases, 36.5% had *grade* marked as unknown in the structured input, but MG_RAG nevertheless assigned a grade. Manual inspection of additional errors showed that MG_RAG sometimes generated a grade different from the input, apparently following the value expressed in the retrieved template. This suggests that retrieval can occasionally override the structured input, either by filling in an unknown value or by introducing a conflicting one.

This effect is substantially less pronounced for Mixtral: Mx_RAG generated an incorrect grade while Mx was correct in only 2.7% of the cases.

Attribute	MedGemma	MedGemma RAG	Mixtral	Mixtral RAG
diag_test	100.00	99.81	100.00	99.72
lymph_nodes_taken	94.63	97.55	91.23	86.90
lymph_nodes_aff	98.77	97.27	90.48	81.81
extracapsular_spread	94.63	92.74	75.97	67.95
tumor_size	99.25	99.25	93.97	93.78
estrogen_receptor	99.25	99.43	98.40	97.55
progest_receptor	100.00	100.00	98.68	97.55
lymphov_invasion	90.01	96.14	73.70	62.77
her2	95.10	97.93	89.35	89.54
clear_margins	85.11	85.30	91.61	83.60
ki67	100.00	100.00	99.25	98.68
morphology	96.32	96.89	97.74	96.23
grade	96.04	94.63	97.17	96.80
Average	**96.08**	**96.69**	**92.12**	**88.68**

Table 7: **Faithfulness accuracy of the Teacher Models on AV_STUDENT.**

Attribute	MedGemma	MedGemma_RAG	Mixtral	Mixtral_RAG
diag_test	99.72	99.91	100.00	99.81
lymph_nodes_taken	95.48	95.01	88.70	91.05
lymph_nodes_aff	99.81	94.07	89.45	86.44
extracapsular_spread	94.92	94.73	88.04	71.85
tumor_size	98.96	99.06	98.31	93.60
estrogen_receptor	100.00	100.00	99.25	98.87
progest_receptor	100.00	100.00	99.72	98.78
lymphov_invasion	96.89	94.54	86.63	69.59
her2	87.48	96.52	85.88	90.87
clear_margins	91.05	87.10	93.97	87.38
ki67	100.00	99.91	99.81	99.06
morphology	96.23	97.74	98.12	97.08
grade	97.08	90.58	98.59	97.36
Average	**96.74**	**96.09**	**94.34**	**90.90**

Table 8: **Faithfulness accuracy of the teacher models on AV_TEACHER.**

The interaction between retrieval and the base model therefore appears to be model-dependent. One possible explanation is that MedGemma relies more strongly on the clinically typical patterns represented in the templates, although this remains a hypothesis.

Downstream impact of filtering. To assess whether the apparent benefit of filtering extends beyond the faithfulness metric used for data selection, we trained downstream information-extraction models on each filtered and unfiltered synthetic dataset and evaluated them on the 377 real, expert-annotated reports.

As shown in Table 9, filtering improves macro-averaged F1 for three of the four teacher conditions: from 0.82 to 0.84 for Mx, from 0.84 to 0.85 for MG_RAG, and from 0.80 to 0.87 for Mx_RAG. For MG, performance remains unchanged at 0.83. Notably, Mx_RAG_Filter achieves the highest downstream F1, although it ranks third in intrinsic faithfulness accuracy. This indicates that the downstream evaluation does not simply reproduce the ranking obtained with the LLM judge.

Teacher condition	Unfiltered F1	Filtered F1	Δ F1
Mx	0.82	0.84	+0.02
MG_RAG	0.84	0.85	+0.01
Mx_RAG	0.80	**0.87**	+0.07
MG	0.83	0.83	0.00

Table 9: Macro-averaged downstream information-extraction F1 of Mistral-7B-Instruct fine-tuned on each filtered and unfiltered synthetic dataset, evaluated on the 377 real, expert-annotated reports.

F Information Extraction Results

Tables 18 and 19 report attribute-level accuracy and F1 scores for the information-extraction models based on MEDGEMMA and Mistral, respectively. We compare the 0-shot and 3-shot baselines with versions of each model fine-tuned on the synthetic datasets.

Attribute	MG	MG_Filt
diag_test	98.77	99.62
lymph_nodes_taken	93.03	95.29
lymph_nodes_aff	99.81	99.72
extracapsular_spread	93.12	94.63
tumor_size	99.81	95.38
estrogen_receptor	100.00	99.62
progest_receptor	100.00	100.00
lymphov_invasion	86.33	92.18
her2	83.13	96.42
clear_margins	84.83	82.66
ki67	99.91	99.91
morphology	94.34	93.87
grade	97.08	98.30
Average	**94.63**	**95.97**

Table 10: Percentage of correctly generated attributes for MG (1062 training examples) and MG_Filt (690 training examples) tested on AV_STUDENT.

Accuracy Bucket	MG	MG_Filt
Perfect (100%)	45.71%	57.02%
High (85–99%)	40.53%	34.68%
Moderate (70–84%)	13.57%	8.29%
Low (50–69%)	0.19%	0.00%
Very Low (< 50%)	0.00%	0.00%

Table 11: Document-level faithfulness accuracy distribution for MG and MG_Filt pathological report generation (AV_STUDENT, N = 1061).

Attribute	MG_RAG	MG_RAG_Filt
diag_test	99.91	99.43
lymph_nodes_taken	97.36	98.59
lymph_nodes_aff	95.48	98.21
extracapsular_spread	90.57	94.25
tumor_size	98.96	99.62
estrogen_receptor	100.00	99.91
progest_receptor	100.00	99.91
lymphov_invasion	92.27	96.14
her2	99.34	98.87
clear_margins	82.94	83.60
ki67	100.00	100.00
morphology	95.29	96.04
grade	75.59	91.99
Average	**94.44**	**96.66**

Table 12: Percentage of correctly generated attributes for MG_RAG (1062 training examples) and MG_RAG_Filt (662 training examples) tested on AV_STUDENT.

Accuracy Bucket	MG_RAG	MG_RAG_Filt
Perfect (100%)	46.18%	64.94%
High (85–99%)	38.64%	28.65%
Moderate (70–84%)	15.08%	5.94%
Low (50–69%)	0.09%	0.47%
Very Low (<50%)	0.00%	0.00%

Table 13: Document-level faithfulness accuracy distribution for MG_RAG and MG_RAG_Filt pathological report generation (AV_STUDENT, N = 1061).

Attribute	Mx	Mx_Filt
diag_test	100.00	100.00
lymph_nodes_taken	87.18	97.27
lymph_nodes_aff	91.42	98.11
extracapsular_spread	87.09	96.42
tumor_size	98.02	99.43
estrogen_receptor	99.34	99.91
progest_receptor	98.49	99.43
lymphov_invasion	67.86	75.87
her2	94.82	96.61
clear_margins	97.27	97.55
ki67	99.81	100.00
morphology	96.98	98.11
grade	98.77	98.11
Average	**93.62**	**96.68**

Table 14: Percentage of correctly generated attributes for Mx (1062 training examples) and Mx_Filt (561 training examples) tested on AV_STUDENT.

Accuracy Bucket	Mx	Mx_Filt
Perfect (100%)	46.09%	63.52%
High (85–99%)	31.86%	30.54%
Moderate (70–84%)	21.30%	5.75%
Low (50–69%)	0.75%	0.19%
Very Low (< 50%)	0.00%	0.00%

Table 15: Document-level faithfulness accuracy distribution for Mx and Mx_Filt pathological report generation (AV_STUDENT, N = 1061).

Attribute	Mx_RAG	Mx_RAG_Filt
diag_test	100.00	100.00
lymph_nodes_taken	92.74	95.57
lymph_nodes_aff	90.95	97.83
extracapsular_spread	80.58	94.63
tumor_size	73.70	98.68
estrogen_receptor	98.96	94.44
progest_receptor	99.72	99.81
lymphov_invasion	68.05	97.83
her2	98.87	98.02
clear_margins	81.24	84.54
ki67	100.00	99.91
morphology	95.76	97.08
grade	96.89	98.02
Average	**90.57**	**96.64**

Table 16: Percentage of correctly generated attributes for Mx_RAG (1062 training examples) and Mx_RAG_Filt (408 training examples) on AV_STUDENT.

Accuracy Bucket	Mx_RAG	Mx_RAG_Filt
Perfect (100%)	33.46%	64.84%
High (85–99%)	29.31%	28.46%
Moderate (70–84%)	31.76%	6.60%
Low (50–69%)	5.47%	0.09%
Very Low (< 50%)	0.00%	0.00%

Table 17: Document-level faithfulness accuracy distribution for Mx_RAG and Mx_RAG_Filt pathological report generation (AV_STUDENT, N = 1061).

	Base Model				Fine-Tuned Models					
	Mg0s		Mg3s		MgS12		MgS123		MgS12_T12	
attribute	acc.	f1	acc.	f1	acc.	f1	acc.	f1	acc.	f1
diagnostic	0.23	0.30	0.45	0.60	0.42	0.55	**0.56**	**0.69**	0.46	0.59
echantillon	**0.89**	**0.93**	0.81	0.88	0.86	0.92	0.85	0.92	0.80	0.89
emboles	0.95	0.95	0.94	0.94	0.95	0.95	0.95	0.96	0.94	0.95
grade	0.88	0.91	0.80	0.85	**0.95**	**0.96**	0.83	0.88	0.90	0.93
her2	0.94	0.95	0.88	0.91	0.94	0.95	0.93	0.95	**0.95**	**0.96**
ki67	0.98	0.98	0.86	0.88	0.98	0.98	0.98	0.98	0.98	0.98
marges_saines	0.86	0.80	0.86	0.80	0.84	0.72	**0.86**	0.79	0.86	0.78
numero_gang_att	0.98	0.97	0.88	0.87	0.98	0.97	0.98	0.97	0.95	0.93
numero_gang_prel	0.98	0.97	0.89	0.87	0.98	0.98	0.96	0.95	0.95	0.94
re	0.93	0.95	0.87	0.90	0.93	0.94	0.91	0.93	**0.97**	**0.98**
rp	0.94	0.95	0.86	0.89	0.93	0.94	0.91	0.93	**0.97**	**0.98**
rupture_capsulaire	0.98	0.87	0.95	0.69	0.98	0.90	0.93	0.36	0.98	0.90
taille	0.58	0.64	0.67	0.68	0.87	0.82	0.85	0.83	0.86	0.81
Total	0.85	0.86	0.82	0.84	**0.89**	**0.90**	0.88	0.89	**0.89**	0.89

Table 18: **Information Extraction Results (MedGemma).** The models fine-tuned on our synthetic data outperform the base model (MedGemma 0- and 3-shots). MgS/T12: MedGemma fine-tuned on combined data from the best two students/teacher models.

	Base Model				Fine-Tuned Models					
	Mi0s		Mi3s		MiS12		MiS123		MiS12_T12	
attribute	acc.	f1	acc.	f1	acc.	f1	acc.	f1	acc.	f1
diagnostic	0.22	0.28	0.16	0.23	0.36	0.52	0.37	0.52	**0.41**	**0.54**
echantillon	0.52	0.68	0.66	0.79	0.68	0.81	0.71	0.83	0.71	0.83
emboles	0.68	0.68	0.73	0.76	0.94	0.94	0.93	0.93	**0.95**	**0.96**
grade	0.55	0.59	0.58	0.66	0.87	0.89	0.87	0.90	**0.88**	0.90
her2	0.49	0.59	0.77	0.81	0.82	0.86	0.92	0.94	0.91	0.93
ki67	0.84	0.84	0.95	0.95	0.96	0.96	0.96	0.96	**0.97**	**0.97**
marges_saines	0.54	0.51	0.66	0.61	0.75	0.39	0.74	0.36	**0.81**	**0.64**
numero_gang_att	0.69	0.43	0.84	0.76	0.95	0.93	0.96	0.95	0.95	0.94
numero_gang_prel	0.89	0.84	0.94	0.93	0.95	0.94	0.97	0.96	0.97	0.96
re	0.59	0.65	0.80	0.84	0.93	0.95	0.93	0.93	**0.94**	0.95
rp	0.56	0.60	0.74	0.77	0.92	0.93	0.85	0.85	**0.95**	**0.95**
rupture_capsulaire	0.54	0.22	0.88	0.52	0.93	0.36	0.96	0.72	0.95	0.62
taille	0.36	0.32	0.34	0.41	0.81	0.75	0.80	0.75	**0.84**	**0.77**
Total	0.57	0.57	0.69	0.71	0.84	0.84	0.84	0.84	**0.86**	**0.86**

Table 19: **Information Extraction Results (Mistral-7B)**. The performance disparity between the base (Mistral-7B 0- and 3-shots) and the models fine-tuned on the synthetic dataset is substantial. MiS/T12: Mistral fine-tuned on combined data from the best two students/teacher models.