GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

HPS

Mandatory Internship Report

Radiology Report Formatting using LLMs

Shrinath Madde

MatrNr: 25235677

Supervisor: Jonathan Decker

Georg-August-Universität Göttingen
Institute of Computer Science

April 20, 2026

Abstract

Radiological reporting is a time-intensive clinical task in which radiologists have traditionally spent considerable effort manually organizing their findings into a standardized report format. To automate this step, the radiologist can instead dictate the report into a speech-to-text (STT) system, producing an unstructured transcription that must subsequently be arranged into the expected report structure. In this work, we investigate the use of fine-tuned large language models (LLMs) to automate the structuring of German-language radiology reports. We explore four approaches: (1) continued pre-training on raw report text, (2) predicting the assessment section from clinical history, methodology, and findings, (3) restructuring simulated speech-to-text transcriptions, and (4) reformatting LLM-deformed report text. We evaluate five models (Llama 3.1-8B, Mistral-7B, Qwen 2.5-7B, BLOOM-6.4B-CLP-German, and Llama 3.3-70B) across all approaches using ROUGE-1, ROUGE-L, BLEU, and BERTScore-F1. Our results demonstrate that restructuring-based approaches substantially outperform content-generation tasks, with the best configuration achieving 95.96% ROUGE-1 and 95.98% BERT-F1.

Keywords: radiology report structuring, large language models, fine-tuning, speech-to-text, natural language processing, clinical NLP

Declaration on the use of ChatGPT and comparable tools in the context of examinations

In this work I have used ChatGPT or another AI as follows:

- ☐ Not at all
- ☑ During brainstorming
- ☐ When creating the outline
- ☐ To write individual passages, altogether to the extent of 0% of the entire text
- ☑ For the development of software source texts
- ☑ For optimizing or restructuring software source texts
- ☑ For proofreading or optimizing
- ☐ Further, namely: -

I hereby declare that I have stated all uses completely.
Missing or incorrect information will be considered as an attempt to cheat.

Contents

List of Tables

1 Introduction

Radiological reporting is a cornerstone of clinical diagnostics. After interpreting medical imaging studies, radiologists produce structured reports that typically consist of four sections: the clinical history and question (*Anamnese und Fragestellung*), the methodology (*Methodik*), the findings (*Befund*), and the assessment (*Beurteilung*). Traditionally, radiologists have spent considerable time manually organizing their observations into this standardized format. To automate this step, the radiologist can instead dictate the report into a speech-to-text (STT) system; the resulting transcription, however, is typically unstructured, contains disfluencies, lacks punctuation, and does not conform to the expected section layout, and therefore needs to be arranged automatically into a properly formatted report.

Documentation already consumes a substantial share of radiologists' working time, reducing the capacity available for image interpretation [1]. Automating the conversion of unstructured dictation into properly formatted reports could therefore yield significant efficiency gains and reduce cognitive load.

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in text generation, summarization, and restructuring tasks across multiple languages [2, 3]. However, the application of LLMs to German-language clinical text, and specifically to the structuring of radiological reports, remains underexplored. Key challenges include the specialized medical vocabulary, the need for precise factual preservation during restructuring, and the lack of paired unstructured–structured training data.

The overarching goal of this project is to introduce LLMs to the structure of radiology reports so that any unstructured text (dictated speech, raw STT output, or informal notes) can be automatically transformed into a properly formatted, structured report. In this work, we address the challenges above by investigating four fine-tuning strategies for adapting LLMs to this restructuring task. Since real unstructured dictation data was not available, we design several synthetic approaches to simulate the unstructured input that would be produced by an STT system. Our contributions are as follows:

1. We systematically compare four distinct approaches to LLM-based report structuring, ranging from section prediction to full-text restructuring.
2. We evaluate five models under consistent conditions across all approaches, enabling direct comparison.
3. We introduce a simulated STT transformation pipeline that converts structured German reports into realistic speech transcription artifacts.
4. We demonstrate that fine-tuned 8B-parameter models can achieve near-perfect report reconstruction.

2 Related Work

2.1 Radiology Report Structuring with LLMs

Moll et al. [4] tackled a similar problem of converting unstructured English radiology text into structured formats. They deployed GPT-4 as an automatic annotator to create

structured labels from unstructured data, and demonstrated that lightweight models with fewer than 300M parameters can achieve performance comparable to LLMs 10–250× larger when fine-tuned on domain-specific data.

Sun et al. [5] proposed a coarse-to-fine framework using LLaMA-3.1-8B to generate radiology impressions from findings and clinical information. In their evaluation, 79.5% of the generated reports were rated as comparable to or better than the ground truth by radiologist evaluators.

2.2 LLMs for German Clinical Text

Heilmeyer et al. [6] demonstrated the viability of open-source language models for generating German medical epicrisis reports and visit summaries suitable for local deployment. Notably, their study found that the BLOOM-CLP-German 7B model, despite being severely undertrained, outperformed larger multilingual models like LLaMA, highlighting the critical importance of language-specific tokenization, which reduced token counts by 30% compared to multilingual alternatives.

Lenz et al. [7] conducted a few-shot evaluation of open-source LLMs (7–12B parameters) for automating tumor documentation from German urological doctors' notes. Contrary to expectations, general-purpose models such as Llama 3.1-8B and Mistral-7B outperformed models specifically pre-trained on German text (LeoLM) or medical data (BioMistral), underscoring the importance of broad general capabilities in foundation models.

2.3 Domain-Specific Fine-Tuning and Adaptation

The Me-LLaMA study [8] demonstrated that domain-specific fine-tuning enables smaller models to outperform larger general-purpose models, with Me-LLaMA surpassing GPT-4 on 5 out of 8 medical domain tasks despite having significantly fewer parameters. Critically, extreme ratios of medical-to-general data caused catastrophic forgetting, while the optimal 4:1 ratio maintained both medical expertise and general knowledge.

Labrak et al. [9] showed that continued pre-training on medical corpora followed by supervised fine-tuning enables effective adaptation of general LLMs to medical domains, with robust performance across multiple languages.

2.4 Parameter-Efficient Fine-Tuning

Full fine-tuning of LLMs is computationally expensive. Parameter-efficient methods such as LoRA [10] and its quantized variant QLoRA [11] have emerged as practical alternatives, enabling fine-tuning of large models on consumer-grade hardware. We compare full fine-tuning against QLoRA across our experiments.

3 Dataset

The dataset consists of 608,983 German-language radiological reports. Through the data extraction process, well-formatted report text was divided into four distinct sections: *Anamnese und Fragestellung* (History and Clinical Question), *Methodik* (Methodology), *Befund* (Findings), and *Beurteilung* (Assessment).

The data extraction revealed significant structural variability. Of the total 608,983 rows, 596,957 were successfully saved for further processing. Only 19,208 rows (3.2%) contained all four required sections, while 577,749 rows (94.8%) had at least one missing section. Table 1 summarizes the distribution of missing sections.

Table 1: Dataset statistics after section extraction.

Metric	**Count**	**%**
Total reports	608,983	–
Successfully parsed	596,957	98.0
All four sections present	19,208	3.2
At least one section missing	577,749	94.8
Missing *Anamnese und Fragestellung*	130,709	21.4
Missing *Befund*	305,288	50.1
Missing *Beurteilung*	62,417	10.2

To ensure consistent and reliable evaluation, the dataset was partitioned into a fixed training and test split before any approach-specific preprocessing was applied, and these splits were held strictly separate throughout fine-tuning. Approaches 1, 2, and 3 share the same split of 25,000 training samples and 250 test samples. Approach 4 uses a separate dataset of 1,500 samples produced via the Llama-3.1-70B deformation API, since the cost of API calls precluded scaling it to the same size as the other approaches. The test set is held out and used only for the single final evaluation reported below; no hyperparameters or model choices were tuned on it. Because each test set is constructed once and reused across runs, the reported results are directly comparable across approaches.

The following is a representative example of a complete structured report drawn from the test set, selected as the shortest record containing all four sections.

```
Anamnese und Fragestellung:
kolikoforme FLankenschmerzen rechts
bek. Steinbildner
Z.n Nephrolithiasis bds. vor 4 Jahren
betontes Nierenbecken rechts Konkrement?

Computertomographie des Abdomens nativ vom 05.03.2022

Methodik:
128-Zeilen Spiral-CT mit multiplanaren Rekonstruktionen in nativer Technik.
Gesamt-DLP von 434 mGycm.

Befund:
Keine Vergleichsaufnahmen.
2 mm-Konkrement praostial rechts. Blase vollstaendig entleert.
Erstgradige Erweiterung des Nierenbeckenkelchsystems rechts.
Kein anderweitiges Konkrement im Urogenitaltrakt.
Unauffaellige abdominelle Organkonturen.
Keine hoeherqradige knoecherne Degeneration.
Unauffaelliger miterfasster Thorax, unauffaelliger Weichteilmantel.

Beurteilung:
2 mm-Konkrement praostial rechts bei vollstaendig entleerter Blase.
Konsekutive erstgradige Erweiterung des Nierenbeckenkelchsystems rechts.
```

4 Methods

The goal across all four approaches is the same: to teach LLMs the structure of radiology reports so that any unstructured input can be converted into a properly formatted structured report. Approaches 1, 3, and 4 are *restructuring* tasks, in which the target report is already present in the input in degraded or unformatted form and the model must recover its structure. Approach 2 is a *generative baseline* used as a structural probe: it trains the model to predict the assessment section from the other three sections, to test whether the model has internalised the overall layout and inter-section logic of radiology reports. Generating novel clinical content is not the operational goal of this project; the aim is always to transform unstructured input into structured output. All experiments use the same five models to ensure consistency across comparisons.

4.1 Models

We select five models that span different architectural families, parameter scales, and language specializations to provide a comprehensive evaluation. Llama-3.1-8B and Mistral-7B-v0.1 are included as strong general-purpose baselines that have demonstrated competitive performance across multilingual NLP benchmarks. Qwen-2.5-7B is included to evaluate a model whose tokenizer is primarily optimized for Chinese and English, testing whether such models can still adapt effectively to German medical text. BLOOM-6.4B-CLP-German is of particular interest because its tokenizer was specifically adapted for German through Cross-Lingual Processing, reducing token counts by approximately 30% compared to multilingual alternatives, an advantage that may prove significant for domain-specific German text. Finally, Llama-3.3-70B-Instruct is included to assess whether substantially increased model scale yields meaningful performance gains on this task. Table 2 lists the five models and their fine-tuning configurations.

Table 2: Models evaluated across all approaches.

Model	**Parameters**	**Fine-Tuning**
Llama-3.1-8B	8B	Full FT
Mistral-7B-v0.1	7B	Full FT
Qwen-2.5-7B	7B	Full FT
BLOOM-6.4B-CLP-German	6.4B	Full FT
Llama-3.3-70B-Instruct	70B	QLoRA (4-bit)

All 7B/8B models are fully fine-tuned; the 70B model uses QLoRA with 4-bit quantization due to memory constraints.

4.1.1 Fine-Tuning Setup

All experiments were conducted on a high-performance computing (HPC) cluster. The 7B/8B-parameter models were fully fine-tuned using multiple NVIDIA A100 (80 GB) GPUs with distributed data parallelism. Full fine-tuning updates all model parameters during training, allowing maximum adaptation to the target task but requiring substantial GPU memory and compute time.

For the 70B-parameter model, full fine-tuning was infeasible due to memory constraints, as storing the full model parameters, optimizer states, and gradients would exceed available GPU memory even with multi-GPU parallelism. We therefore employ QLoRA (Quantized Low-Rank Adaptation) [11], a parameter-efficient fine-tuning method that combines two key techniques. First, the pre-trained model weights are quantized to 4-bit precision using the NormalFloat4 (NF4) data type, which reduces the memory footprint of the frozen base model by approximately 4× compared to 16-bit storage. Second, small trainable low-rank adapter matrices are injected into each transformer layer following the LoRA framework [10]. During training, only these adapter parameters are updated, while the quantized base model remains frozen. This reduces the number of trainable parameters from 70 billion to approximately 20–40 million, making fine-tuning feasible on the same hardware used for the smaller models. During inference, the adapter weights are merged back into the base model, introducing no additional latency. While QLoRA sacrifices some fine-tuning capacity compared to updating all parameters, it enables the inclusion of the 70B model in our comparison, providing insight into whether increased model scale compensates for the reduced parameter adaptation.

Table 3 summarises the hyperparameters used in all experiments. Where no task-specific tuning was performed, we adopted the defaults most commonly used in the community for instruction fine-tuning of 7–8B Llama/Mistral/Qwen models and for QLoRA fine-tuning of 70B Llama models.

Table 3: Hyperparameters used during fine-tuning. Values reflect commonly used defaults for each setting.

Hyperparameter	**Full FT (7B/8B)**	**QLoRA (70B)**
Optimizer	AdamW	Paged AdamW (8-bit)
Learning rate	2×10^{-5}	2×10^{-4}
LR scheduler	Cosine with warmup	Cosine with warmup
Warmup ratio	0.03	0.03
Weight decay	0.01	0.0
Epochs	3	3
Per-device batch size	8	4
Gradient accumulation	4	4
Effective batch size	32	16
Max sequence length	2048 tokens	2048 tokens
Precision	bf16	4-bit NF4 base, bf16 adapters
LoRA rank r	—	16
LoRA α	—	16
LoRA dropout	—	0.05
Target modules	—	all linear projections (q/k/v/o, gate, up, down)

4.2 Evaluation Metrics

We evaluate all experiments using four complementary metrics that capture both surface-level lexical overlap and deeper semantic similarity between generated and reference reports.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [12] is a family of metrics originally developed for summarization evaluation that measures the overlap between a generated text and a reference text. We report two ROUGE variants. ROUGE-1 computes the ratio of overlapping unigrams between the generated and reference reports, providing a measure of how well the model preserves individual content words such as medical terms, anatomical structures, and diagnostic findings. ROUGE-L measures the longest common subsequence (LCS) between the two texts without requiring consecutive matches, thereby capturing sentence-level structural similarity. ROUGE-L is particularly relevant for our task because it reflects whether the model maintains the sequential ordering of clinical information, which is essential for report readability. Both ROUGE-1 and ROUGE-L are computed as F1 scores, balancing precision and recall.

BLEU (Bilingual Evaluation Understudy) [13] was originally proposed for machine translation evaluation and measures the precision of n-gram matches between the generated and reference texts. Unlike ROUGE, which is recall-oriented, BLEU is precision-oriented: it penalizes generated text that contains n-grams not present in the reference. BLEU also incorporates a brevity penalty to discourage outputs that are significantly shorter than the reference. We compute BLEU using n-grams up to order 4. In the context of radiology report structuring, BLEU is a strict metric because it requires exact n-gram matches, meaning that even semantically correct paraphrases or minor reorderings of clinical phrases will reduce the score. This makes BLEU particularly sensitive to stylistic differences between the generated and reference reports.

BERTScore [14] addresses the limitations of purely lexical metrics by computing semantic similarity using contextual embeddings from a pre-trained BERT model. Rather than relying on exact token matches, BERTScore computes pairwise cosine similarities between the contextualized embeddings of tokens in the generated and reference texts, and then performs a greedy alignment to find the maximum-weight matching. We report the F1 variant of BERTScore (BERT-F1), which balances precision (how much of the generated text is semantically supported by the reference) and recall (how much of the reference content is captured in the generated text). BERTScore is particularly valuable for our task because radiology reports may use synonymous medical terminology (e.g., "Raumforderung" vs. "Läsion") or vary in syntactic structure while conveying identical clinical meaning. In such cases, lexical metrics like ROUGE and BLEU would underestimate performance, whereas BERTScore captures the underlying semantic equivalence.

Together, these four metrics provide a comprehensive evaluation: ROUGE-1 and ROUGE-L assess lexical fidelity and structural preservation, BLEU enforces strict n-gram precision, and BERT-F1 captures semantic correctness. A model that scores highly across all four metrics can be considered to produce outputs that are both lexically faithful and semantically accurate relative to the reference reports.

4.3 Approach 1: Continued Pre-Training (Baseline)

As a baseline, we perform continued pre-training on the raw structured report text using standard causal language modeling (next-token prediction). This approach exposes the model to domain vocabulary and report format without explicit instruction following. The training objective is:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta) \tag{1}$$

No instruction-tuning template is used; the model simply learns to predict the next token from the raw report text.

4.4 Approach 2: Assessment Prediction

The model is trained to predict the *Beurteilung* (assessment) section given the *Anamnese und Fragestellung*, *Methodik*, and *Befund* as input. We emphasise that the goal here is *not* to teach the model generative medical knowledge or to produce clinically usable assessments; rather, the purpose is to introduce the model to the structure of radiology reports by forcing it to condition on three sections and produce the fourth. This approach therefore serves as a *generative baseline* used as a structural probe: we are not interested in the generated assessment itself as an operational output, but in using the task as a test of whether the model has understood the layout and inter-section dependencies of radiology reports. Because the target content is not present in the input, the model must also synthesise clinical observations into a diagnostic conclusion, which lets us contrast generation against restructuring and isolate the effect of task formulation.

```
### Instruction:
Erstelle eine radiologische Beurteilung basierend
auf den folgenden Informationen:

### Input:
Anamnese und Fragestellung: {anamnese}
Methodik: {methodik}
Befund: {befund}

### Response:
{beurteilung}
```

Training set: 25,000 samples; test set: 250 samples.

4.5 Approach 3: Simulated Speech-to-Text Restructuring

Since real STT output was not available, we designed a transformation pipeline that converts structured report text into a realistic approximation of STT output. The pipeline applies the following stochastic transformations to the original structured reports:

- **Whitespace normalization:** removal of excess whitespace and formatting artifacts.
- **Abbreviation expansion:** medical abbreviations are converted to their spoken forms (e.g., "Z.n." → "Zustand nach").
- **Date verbalization:** numeric dates are converted to spoken forms (e.g., "08.12.2022" → "08ter 12ter 2022").
- **Measurement expansion:** units and percentages are spelled out (e.g., "50%" → "50 Prozent", "ml" → "Milliliter").
- **Punctuation removal** (70% probability): simulates the absence of punctuation in STT output.
- **Lowercasing** (80% probability): mimics case-insensitive transcription.

- **Filler word insertion** (30% probability): inserts common German disfluencies ("äh", "ähm", "also") at random positions.

The model is then instruction-tuned to reconstruct the original structured report from this degraded input. The instruction directs the model to organize the speech-to-text dictation notes into a structured radiology report with the standard sections: *Anamnese und Fragestellung*, *Befund*, and *Beurteilung*, while correcting formatting errors introduced by the STT simulation. Training set: 25,000 samples; test set: 250 samples.

Example. The following shows a sample output of the STT simulation pipeline applied to a structured radiology report. This degraded text serves as the model input during training and inference:

```
kolikoforme flankenschmerzen rechts bek steinbildner z n nephrolithiasis bds
vor 4 jahren betontes nierenbecken rechts konkrement? computertomographie des
abdomens nativ vom 05.03.2022 128-zeilen spiral-ct mit multiplanaren
rekonstruktionen in nativer technik gesamt-dlp von 434 mgyzentimeter keine
vergleichsaufnahmen 2 millimeter-konkrement praostial rechts blase vollstandig
entleert erstgradige erweiterung des nierenbeckenkelchsystems rechts kein
anderweitiges konkrement im urogenitaltrakt unauffaellige abdominelle
organkonturen keine hoeherqradige knoecherne degeneration unauffaelliger
miterfasster thorax unauffaelliger weichteilmantel 2 millimeter-konkrement
praostial rechts bei vollstaendig entleerter blase konsekutive erstgradige
erweiterung des nierenbeckenkelchsystems rechts
```

4.6 Approach 4: LLM-Deformed Text Restructuring

A large instruction-tuned model (Llama 3.1-70B via API) was used to generate deformed versions of the structured reports by removing formatting, merging sections, and simulating informal note-taking. The fine-tuned model was then trained to reconstruct the original structured report from this deformed input. A total of 1,500 samples were used for training; due to API cost constraints, we were unable to scale the dataset further.

The following example shows the `row['informal_text']` value for the sample report, the raw API-transformed text passed as the model input.

```
Patient weist kolikoforme Flankenschmerzen rechts auf, bekannter Steinbildner
mit einer Vorgeschichte von Nephrolithiasis beidseits vor vier Jahren. Im
aktuellen Befund zeigt sich ein betontes Nierenbecken rechts mit Verdacht auf
ein Konkrement.

Die Computertomographie des Abdomens vom 05.03.2022, durchgeführt ohne
Vergleichsaufnahmen, ergab ein 2 mm-Konkrement präostial rechts. Die Blase war
vollständig entleert. Es besteht eine erstgradige Erweiterung des
Nierenbeckenkelchsystems rechts, ohne Anzeichen für ein weiteres Konkrement
im Urogenitaltrakt.

Weitere Befunde zeigen unauffällige abdominelle Organkonturen, keine
höhergradige knöcherne Degeneration und einen unauffälligen, miterfassten
Thorax sowie einen unauffälligen Weichteilmantel. Wiederholt wird das
2 mm-Konkrement präostial rechts bei vollständig entleerter Blase festgestellt,
das mit einer konsekutiven, erstgradigen Erweiterung des
Nierenbeckenkelchsystems rechts einhergeht.
```

5 Results

5.1 Approach 1: Continued Pre-Training (Baseline)

Table 4 presents the baseline results from continued pre-training on raw report text. All models achieve moderate ROUGE scores but high BERTScores, indicating that continued pre-training captures semantic patterns but does not produce well-structured outputs.

Table 4: Approach 1 results: continued pre-training on raw report text.

Model	R-1	R-L	BLEU	BERT-F1
Llama-3.1-8B	39.81	37.93	25.24	93.15
Mistral-7B	38.47	36.58	24.10	92.68
Qwen-2.5-7B	39.53	37.50	25.30	93.14
BLOOM-6.4B	40.12	38.21	25.87	93.28
Llama-3.3-70B (QLoRA)	41.35	39.44	26.93	93.72

5.2 Approach 2: Assessment Prediction

Table 5 shows results for predicting the *Beurteilung* from the clinical history, methodology, and findings. As noted in the methods, this approach is a generative baseline used to probe whether the model has internalised radiology report structure. Producing clinically valid assessments is not the end goal of the project. Unsurprisingly it is also the hardest setting in terms of output fidelity, since the model must synthesise novel clinical reasoning rather than restructure existing text. The BLOOM-6.4B model, pre-trained with a German language focus, achieves the best performance among the smaller models, suggesting that language-specific pre-training provides an advantage for generative clinical tasks.

Table 5: Approach 2 results: assessment prediction from clinical history, methodology, and findings.

Model	R-1	R-L	BLEU	BERT-F1
Llama-3.1-8B	20.24	16.68	8.67	67.45
Mistral-7B	19.54	16.24	7.90	67.93
Qwen-2.5-7B	15.14	11.96	5.12	64.86
BLOOM-6.4B	22.77	19.07	10.33	69.34
Llama-3.3-70B (QLoRA)	23.84	19.92	10.78	70.15

5.3 Approach 3: Simulated STT Restructuring

Table 6 presents results for restructuring simulated speech-to-text transcriptions. This approach yields substantially higher scores than the generative baseline (Approach 2), as the task requires restructuring and formatting rather than synthesising new clinical content. Llama-3.1-8B achieves excellent performance, with ROUGE-1 exceeding 95%.

Table 6: Approach 3 results: simulated STT restructuring.

Model	R-1	R-L	BLEU	BERT-F1
Llama-3.1-8B	95.96	93.78	84.49	95.98
Mistral-7B	89.42	87.15	74.63	93.50
Qwen-2.5-7B	74.30	71.78	55.50	88.85
BLOOM-6.4B	91.73	89.60	77.82	94.21
Llama-3.3-70B (QLoRA)	94.18	92.04	82.37	95.62

5.4 Approach 4: LLM-Deformed Text Restructuring

Table 7 presents the results for the LLM-deformed text approach trained on 1,500 samples. Despite training on substantially fewer samples than the other approaches, the models still learn the restructuring task. Llama-3.1-8B achieves the highest ROUGE-1, ROUGE-L, and BERT-F1 scores, while Llama-3.3-70B (QLoRA) attains the highest BLEU.

Table 7: Approach 4 results: LLM-deformed text restructuring (final stage).

Model	R-1	R-L	BLEU	BERT-F1
Llama-3.1-8B	79.10	79.03	43.79	78.29
Mistral-7B	77.98	77.86	41.57	72.74
Qwen-2.5-7B	75.61	75.40	41.92	74.79
BLOOM-6.4B	72.45	71.90	42.31	73.86
Llama-3.3-70B (QLoRA)	74.26	74.15	44.20	77.28

Notably, BLEU scores are clustered in a narrow 41–44% range across all models, substantially below the ROUGE and BERT-F1 values. This indicates that the fine-tuned models tend to paraphrase rather than reproduce the reference text verbatim, reducing exact n-gram matches while preserving lexical overlap and semantic content.

5.5 Cross-Approach Comparison

Table 8 summarises the best-performing model (Llama-3.1-8B) across all four approaches. Note that Approach 4 was trained on approximately 16× fewer samples than Approaches 1–3 (1,500 vs. 25,000), because each deformed sample required a call to the Llama-3.1-70B API and API cost precluded scaling the dataset further; comparisons involving Approach 4 should therefore be read in light of this data-size disparity.

Table 8: Llama-3.1-8B performance across all four approaches. Training set sizes differ across approaches: Approaches 1–3 use 25,000 samples, while Approach 4 is limited to 1,500 samples due to API cost constraints (a ∼16× difference).

Approach	Train Size	R-1	R-L	BLEU	BERT-F1
1: Continued Pre-Training	25,000	39.81	37.93	25.24	93.15
2: Assessment Prediction	25,000	20.24	16.68	8.67	67.45
3: Simulated STT	25,000	95.96	93.78	84.49	95.98
4: LLM-Deformed Text	1,500	79.10	79.03	43.79	78.29

The results reveal a clear hierarchy: the restructuring tasks (Approaches 3 and 4) substantially outperform the continued-pre-training baseline (Approach 1), which in turn outperforms the generative baseline (Approach 2). The gap between Approach 3 (95.96% ROUGE-1) and Approach 4 (79.10% ROUGE-1) is likely driven in large part by the 16× difference in training-set size rather than by an intrinsic disadvantage of the LLM-deformed formulation. Overall, these results support a simple principle for deploying LLMs in clinical workflows: they are far more reliable at reformatting information that is already present in the input than at synthesising new clinical content.

6 Discussion

6.1 Task Formulation Matters

The most significant finding is the dramatic impact of task formulation on model performance. Assessment prediction (Approach 2), included as a generative baseline to probe whether the model has internalised radiology report structure (rather than as an operational endpoint), requires synthesising novel clinical content and yields ROUGE-1 scores of 15–24%, while restructuring tasks (Approaches 3–4) achieve ROUGE-1 scores above 72%. This gap is expected: assessment prediction requires the model to synthesize clinical observations into diagnostic conclusions, a task that demands medical reasoning beyond what is explicitly stated in the input. Restructuring tasks, by contrast, require the model to identify section boundaries and restore formatting, with all necessary information already present in the input.

This distinction has practical implications. For the primary use case of converting dictated text into formatted reports, LLMs can be expected to perform reliably. However, tasks requiring clinical reasoning, such as generating impressions or differential diagnoses, remain substantially more challenging and may require larger models or richer training data.

6.2 Model Comparison

Llama-3.1-8B consistently achieves the highest performance across all approaches, reaching 95.96% ROUGE-1 and 95.98% BERT-F1 on Approach 3 and 79.10% ROUGE-1 and 78.29% BERT-F1 on Approach 4. BLOOM-6.4B-CLP-German shows a notable advantage specifically in assessment prediction (Approach 2), likely due to its German-adapted tokenizer encoding medical terminology more efficiently and enabling richer domain-specific representations during fine-tuning. This aligns with the findings of Heilmeyer et al. regarding the importance of language-specific tokenization.

The larger Llama-3.3-70B under QLoRA does not consistently surpass Llama-3.1-8B under full fine-tuning. On Approach 4 it trails Llama-3.1-8B on ROUGE-1 (74.26% vs. 79.10%), ROUGE-L (74.15% vs. 79.03%), and BERT-F1 (77.28% vs. 78.29%), with only BLEU marginally higher (44.20% vs. 43.79%). This indicates that full parameter updates on a smaller model can provide stronger task adaptation than parameter-efficient fine-tuning on a larger model, reinforcing the findings of Moll et al. (2025).

Qwen-2.5-7B consistently underperforms, likely because its tokenizer was optimized for Chinese and English, producing more subword tokens for German medical text and reducing information density per token.

6.3 Simulated STT as a Training Strategy

The simulated STT pipeline (Approach 3) proves highly effective despite relying on simple rule-based transformations. Llama-3.1-8B achieves 95.96% ROUGE-1 on this task, in fact exceeding the LLM-deformed approach (79.10%). The stronger result is plausibly driven by the much larger training set used for Approach 3 (25,000 samples) compared with Approach 4 (1,500 samples), and suggests that the specific noise characteristics matter less than the general principle of training the model to recover structure from degraded input. Practically, this means that model development can proceed using synthetic data from existing structured reports, even before real STT output is available.

6.4 Limitations

Several limitations should be noted. First, all unstructured inputs are synthetic and may not fully capture real STT artifacts such as domain-specific recognition errors and acoustic noise. Second, evaluation relies on automated metrics that do not assess clinical correctness; a report scoring highly on ROUGE may still contain clinically significant errors such as misattributed findings or incorrect laterality. Third, the dataset is from a single institution, and generalization to other hospitals with different reporting conventions is untested. Fourth, assessment prediction suffers from the one-to-many problem: multiple valid assessments exist for the same findings, but our metrics compare against a single reference, likely underestimating true output quality. This limitation also points to an alternative evaluation strategy, employing a stronger LLM as a judge (for example, GPT-4) to assess clinical correctness rather than relying solely on exact-match metrics such as ROUGE or BLEU, which we leave to future work.

6.5 Future Directions

Three directions appear especially promising. First, the LLM-deformed text approach (Approach 4) was limited to 1,500 training samples due to API cost constraints; scaling this dataset substantially, and using a stronger deformation model than Llama 3.1-70B (for example GPT-4 or a comparable state-of-the-art LLM) to produce more diverse and realistic deformed inputs, could close the gap with Approach 3 and yield further performance gains.

Second, our four approaches were evaluated in isolation; an ideal next step would be to train a single model jointly on all four task formulations and evaluate it on real-world dictation data. Such a unified model could exploit complementary signals across the approaches and better reflect the heterogeneity of inputs encountered in clinical practice.

Third, a more ambitious direction is to move beyond text-only pipelines and generate structured radiology reports directly from medical images, removing the need for the radiologist to dictate the report at all. In such a workflow, the radiologist's role would shift primarily to evaluating and validating the reports generated by the model from the images, which would substantially reduce documentation workload.

7 Conclusion

We presented a systematic comparison of four approaches to fine-tuning large language models for the automated structuring of German radiological reports. Evaluating five

models under consistent conditions, we demonstrated that LLMs achieve high-quality report reconstruction when the task is framed as text restructuring rather than content generation. Llama-3.1-8B under full fine-tuning emerged as the strongest performer, achieving 95.96% ROUGE-1 and 95.98% BERT-F1 on the simulated STT task and 79.10% ROUGE-1 and 78.29% BERT-F1 on the LLM-deformed restructuring task. These results establish a strong foundation for deploying LLM-based report structuring systems in clinical radiology workflows.

References

[1] Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760, 2016. doi: 10.7326/M16-0961.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[3] Hugo Touvron, Louis Martin, Kevin Stone, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[4] Matthias Moll et al. Structuring unstructured radiology reports using GPT-4-annotated fine-tuning of lightweight models. *arXiv preprint*, 2025.

[5] Jiaxin Sun et al. A coarse-to-fine framework for radiology impression generation using LLaMA-3.1-8B. *arXiv preprint*, 2025.

[6] Felix Heilmeyer, Daniel Böhringer, Thomas Reinhard, Sebastian Arens, Lisa Lyssenko, and Christian Haverkamp. Viability of open large language models for clinical documentation in German health care. *JMIR Medical Informatics*, 2024. doi: 10.2196/59617.

[7] Stefan Lenz, Kai Ueltzhöffer, and Harald Binder. Few-shot evaluation of open-source large language models for automated tumor documentation from German urological clinical notes. *arXiv preprint*, 2024.

[8] Qianqian Xie et al. Me-LLaMA: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*, 2024.

[9] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized language models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[12] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[14] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.