Semantic retrieval-augmented translation for radiology reports using small language models: Algorithm development and evaluation

Semantische Retrieval-gestützte Übersetzung von Radiologiebefunden mit kleinen Sprachmodellen: Algorithmusentwicklung und -evaluation

Daniel Reichenpfader[1,2]
Fabio Dennstädt[3]

1 Faculty of Medicine, University of Geneva, Switzerland
2 Institute for Patient-Centered Digital Health, Bern University of Applied Sciences, Bern, Switzerland
3 Department of Radiation Oncology, Bern University Hospital, University of Bern, Switzerland

Abstract

Introduction: Accurate machine translation (MT) of clinical text is essential for multilingual patient care, cross-border data sharing, and research reproducibility, yet the use of commercial APIs for such tasks raises substantial privacy concerns. We present Semantic Retrieval-Augmented Translation (S-RAT), a lightweight, ontology-guided framework that integrates the Unified Medical Language System (UMLS) and SNOMED CT concept retrieval into large language model prompts to improve medical terminology fidelity without fine-tuning or external APIs.
Methods: Using 500 English radiology reports from MIMIC-CXR, we evaluated multiple S-RAT variants with open-source language models (gemma3, med42-v2, deepseek-r1, and gpt-oss) as well as a commercial frontier model (Gemini 3 Pro). We compared reference-free quality estimation metrics (COMET-src and GEMBA-DA-noref) with human annotation. Furthermore, we assessed the translation coverage of the German SNOMED CT edition for the specific use case of radiology report translation.
Results: S-RAT prompting did not consistently outperform standard prompting, and COMET-src correlated only weakly ($\rho \approx 0.14$, $p \approx 0.05$) with human adequacy judgments, while GEMBA-DA-noref showed no meaningful correlation. The German SNOMED CT edition covered only 32.6% of extracted concepts.
Discussion: These findings highlight the current limitations of multilingual ontology resources and the application of general-domain MT metrics in clinical contexts. S-RAT demonstrates the feasibility of privacy-preserving, locally deployable, and ontology-augmented translation and provides a foundation for future work on domain-specific evaluation, ontology enrichment, and local deployment of language models in healthcare settings.

Keywords: machine translation, few-shot/zero-shot MT, human evaluation, domain adaptation, healthcare applications, natural language processing, clinical NLP, NLP in resource-constrained settings

Zusammenfassung

Einleitung: Eine präzise maschinelle Übersetzung (MT) klinischer Texte ist entscheidend für die mehrsprachige Patientenversorgung, den grenzüberschreitenden Datenaustausch und die Reproduzier-barkeit von Forschung. Der Einsatz kommerzieller APIs für solche Anwendungen wirft jedoch erhebliche Datenschutzbedenken auf. Wir stellen Semantic Retrieval-Augmented Translation (S-RAT) vor, ein leichtgewichtiges, ontologiegestütztes Framework, das die Konzeptextraktion aus dem Unified Medical Language System (UMLS) und SNOMED CT in Prompts für große Sprachmodelle integriert, um die terminologische Genauigkeit ohne Fine-Tuning oder externe APIs zu verbessern.

Methoden: Auf Basis von 500 englischsprachigen radiologischen Befunden aus MIMIC-CXR evaluierten wir mehrere S-RAT-Varianten unter Verwendung quelloffener Sprachmodelle (gemma3, med42-v2, deepseek-r1 und gpt-oss) sowie eines kommerziellen Frontier-Modells (Gemini 3 Pro). Wir verglichen referenzfreie Qualitätsmetriken (COMET-src und GEMBA-DA-noref) mit menschlichen Annotationen. Zusätzlich untersuchten wir die Abdeckung der deutschen SNOMED-CT-Edition für den spezifischen Anwendungsfall der Übersetzung radiologischer Befunde.
Ergebnisse: S-RAT-Prompting zeigte keinen konsistenten Vorteil gegenüber Standard-Prompting. COMET-src korrelierte nur schwach mit menschlichen Bewertungen der inhaltlichen Angemessenheit ($\rho \approx 0{,}14$, $p \approx 0{,}05$), während GEMBA-DA-noref keine relevante Korrelation aufwies. Die deutsche SNOMED-CT-Edition deckte lediglich 32,6% der extrahierten Konzepte ab.
Diskussion: Die Ergebnisse verdeutlichen die aktuellen Einschränkungen mehrsprachiger Ontologieressourcen sowie die begrenzte Eignung allgemeiner MT-Evaluationsmetriken im klinischen Kontext. S-RAT demonstriert die Machbarkeit einer datenschutzkonformen, lokal einsetzbaren und ontologiegestützten Übersetzung und bildet eine Grundlage für zukünftige Arbeiten zur domänenspezifischen Evaluation, zur Erweiterung von Ontologien sowie zum lokalen Einsatz von Sprachmodellen im Gesundheitswesen.

Schlüsselwörter: maschinelle Few-Shot-/Zero-Shot-Übersetzung, Evaluation, natürliche Sprachverarbeitung, Anwendungen im Gesundheitswesen, Computerlinguistik

1 Introduction

Large language models (LLMs) have demonstrated strong capabilities in few-shot and zero-shot machine translation (MT), including in the biomedical domain [1]. In clinical practice, accurate MT of free-text documents such as radiology reports is essential for cross-lingual collaboration, international patient care, and dataset interoperability. In research, multilingual datasets are necessary for comparing the performance of developed algorithms across languages. However, in many countries, regulatory and privacy concerns prevent the submission of protected health information to commercial APIs, thereby limiting the applicability of leading proprietary LLMs in real-world clinical use cases. Open-source models such as gemma3 [2], gpt-oss [3], and Apertus [4] have been rapidly improving in performance, and for single-turn translation tasks, model size is less critical than in dialogue or reasoning-heavy settings. This creates an opportunity for smaller, locally deployable models (<15 billion parameters) in hospital infrastructures, but also raises questions about how effectively such models can be guided using structured external knowledge.

Existing biomedical MT approaches rely on either large proprietary models, which raise privacy concerns, or open models fine-tuned with large annotated corpora. However, neither approach leverages structured clinical ontologies for translation guidance. We propose a method called semantic retrieval-augmented translation (S-RAT), which identifies Unified Medical Language System (UMLS) [5] concepts in the source text, retrieves the corresponding SNOMED CT preferred terms in the target language, and injects these translations directly into the prompt. This aims at guiding the LLM in accurate translation of clinical terms without requiring fine-tuning or cloud APIs. Unlike standard retrieval-augmented generation (RAG), which retrieves unstructured text, S-RAT performs structured retrieval from the SNOMED CT ontology, mapping source-language medical entities to standardized target-language terms. This structured retrieval reduces ambiguity and ensures terminological consistency, particularly important in clinical subdomains such as radiology.

S-RAT is designed to be lightweight, privacy-preserving, and reproducible. With this paper, we evaluate whether the German national SNOMED CT edition contains sufficient translated concepts for the radiology domain, and whether retrieval-augmented prompting affects translation accuracy compared with standard prompting. We further assess whether COMET-src, a reference-free MT quality estimation metric, aligns with human expert judgments. Specifically, we test the following hypotheses:

- H1: Translation coverage exceeds 80%, defined as the proportion of extracted concept instances for which a German translation could be retrieved.
- H2: Retrieval-augmented translation with S-RAT using a small open-source LLM (gemma3:4b) differs from standard prompting.
- H3: The two general-domain quality estimation metrics COMET-src and GEMBA-DA-noref correlate at most weakly ($\rho < 0.3$) with human judgment on radiology report translations.

1.1 Related work

In the biomedical domain, both MT and LLMs have been evaluated extensively. A recent study comparing GPT-4, GPT-3.5 and Qwen1.5 found that GPT-4 provided the most accurate translations of CT and MRI reports across nine languages [6]. Similarly, a comparative analysis of ChatGPT (GPT-4) and Google Translate on patient discharge instructions reported higher accuracy of ChatGPT [1]. To support such evaluations and facilitate the traceability of results, various multilingual corpora have been introduced. The MIMIC-CXR dataset provides large-scale radiology reports but only in English. The OpenWHO dataset comprises 26,824 sentences across more than 20 languages [7], the Multilingual Medical Corpus offers a large collection of biomedical texts across English, Spanish, French, and Italian [8], and MedEV contains approximately 360,000 Vietnamese–English sentence pairs for medical MT [9]. Despite these advances, the availability of open multilingual clinical datasets, especially for radiology, remains limited.

Beyond datasets, several strategies have been proposed to adapt LLMs for medical MT. Rios fine-tuned LLMs with medical glossaries via QLoRA, achieving improved MT evaluation scores with higher terminology accuracy [10]. RAG has also emerged as a promising technique to enhance translation by injecting relevant knowledge at inference time [11]. RAGtrans introduced a benchmark of 79,000 knowledge-intensive sentences and demonstrated that retrieval-augmented translation with unstructured documents can outperform instruction-tuning alone [12].

In contrast to these approaches, our method does not rely on large fine-tuning datasets or unstructured document retrieval. Instead, we leverage structured clinical ontologies: SNOMED CT concept translations are directly retrieved and incorporated into prompts, upgrading a small, locally deployed LLM with improved terminology coverage in a privacy-compliant way.

2 Methods

In this section, we provide an overview of the complete S-RAT pipeline as well as introduce the applied datasets, models, tools, and evaluation metrics.

2.1 S-RAT pipeline

To augment LLM prompts with quality-assured translations of extracted medical terms, we implement a translation retrieval pipeline: Figure 1 summarizes the complete S-RAT workflow. Starting from an English radiology report, clinical concepts are extracted and normalized to SNOMED CT identifiers, translated using the German SNOMED CT edition (or a fallback dictionary if necessary), and finally injected into the translation prompt before the LLM generates the German report. The figure illustrates where structured terminology retrieval augments an otherwise standard prompting workflow. Licenses and tools are detailed in Appendix A (Attachment 1).

1. **Concept extraction:** Clinical concepts are identified and extracted using a locally deployed commercial service (Azure Text Analytics for health) [13].
2. **Normalization:** UMLS Concept Unique Identifiers (CUIs) and, if available, the SNOMED CT identifier (SCTID) are obtained together with the extracted concepts. If the SCTID is not provided, it is obtained via the UMLS REST API [14].
3. **Multilingual translation:** Verified translations of the identified concepts are retrieved from a locally deployed SNOMED CT terminology server (Snowstorm), which is populated with the respective national edition of SNOMED CT (in this case, German) [15].
4. **Fallback translation:** If a translation is not found in the national edition of SNOMED CT, the system attempts to retrieve the translation from a manually populated list of human-verified translations of common concepts.
5. **Prompt construction:** The translated terms are included within the user prompt.

2.2 Datasets, models, and prompts

We evaluate our method using the MIMIC-CXR dataset, which contains 227,835 imaging studies including reports for 65,379 patients presenting to the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016 [16]. For our analysis, we randomly sampled 500 reports. In addition, we selected two reports manually for in-context learning. Details on data preprocessing and sample characteristics are provided in Appendix B (Attachment 1).

Our experiments are conducted with open-source models that can be realistically deployed locally on institutional infrastructure, while still being large enough to perform MT. Therefore, we chose model sizes between four and twelve billion parameters and focused on the gemma3 family, a suite of state-of-the-art, medium-sized open-source generative models with strong multilingual capabilities [2]. The smallest 1b variant was excluded after preliminary tests showed poor overall performance, which resulted in using the 4b and 12b model variants. We also include med42-v2:8b, a fine-tuned model for the medical domain, based on llama3.3 [17] as well as deepseek-r1:8b [18]. As state-of-the-art baseline, we use gpt-oss:120b, the strongest open-source model available at the time of writing [3]. Additionally, we use Gemini 3 Pro as state-of-the-art frontier model [19].

We base our initial system and user prompts on the work of Chen et al. [20], who evaluated various MT tools for translating critical care-related content, building on an earlier industry report. We adapt their prompt to the use case of radiology report translation. To enhance the concept translation module, we further developed a more sophisticated system prompt. This refined version incorporates additional instructions and leverages a two-shot

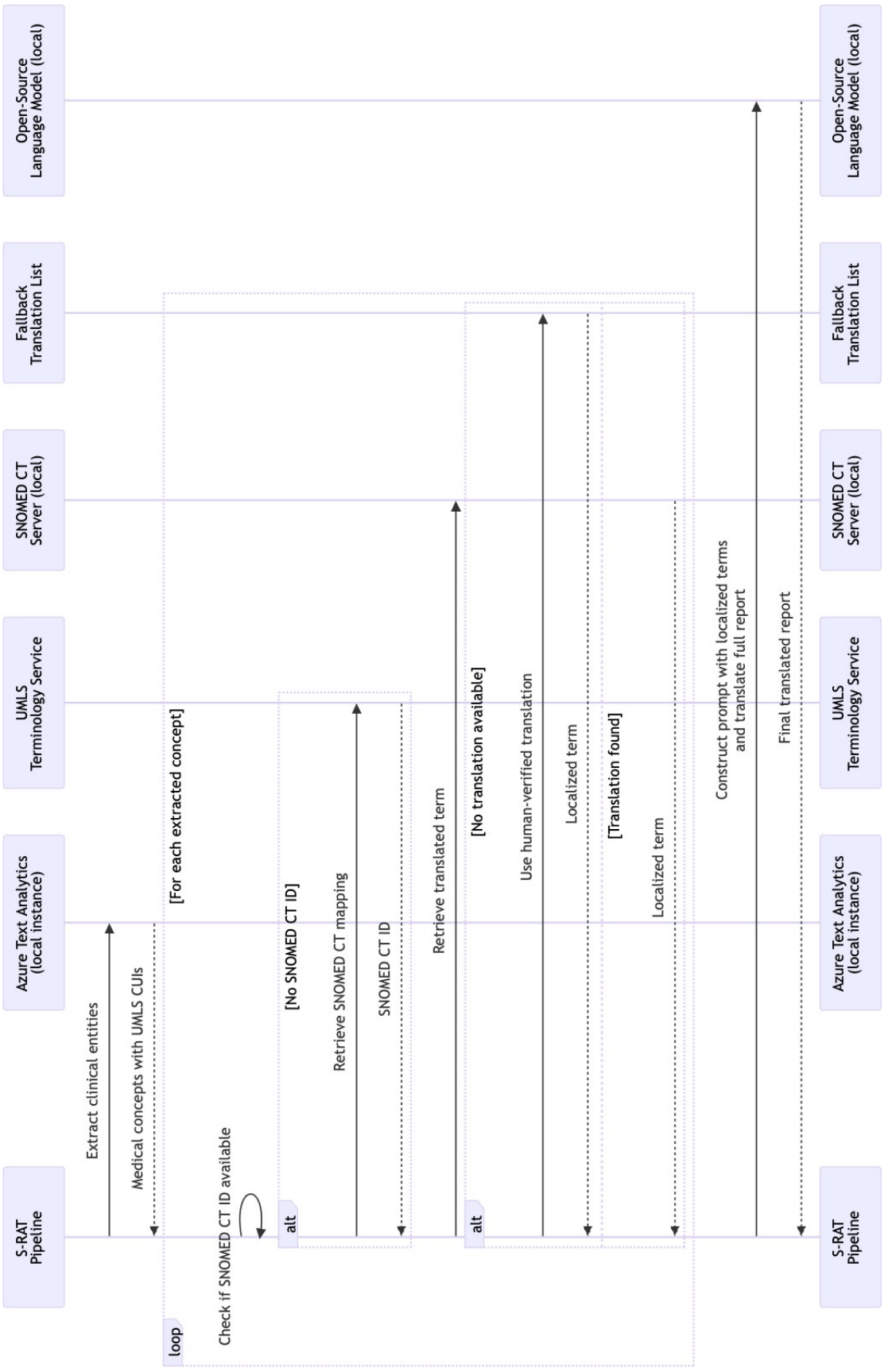

Figure 1: Overview of the S-RAT retrieval pipeline. Clinical concepts are extracted from the source report, mapped to SNOMED CT concepts, translated using the German SNOMED CT edition (with optional fallback dictionary), and inserted into the prompt to guide the LLM during translation.

in-context learning setup with two manually translated reports, used in the S-RAT-Shot and S-RAT-Curated variants. Prior work by Brown et al. has shown that two examples can already substantially improve performance [21]. The full set of prompts is provided in Appendix G (Attachment 1). All resources will be made publicly available via Zenodo [22].

2.3 Evaluation metrics

MT evaluation has traditionally been relying on reference-based metrics, which compare system outputs against human-produced reference translations. Examples for reference-based metrics are BLEU [23], ROUGE [24], METEOR [25], and BERTScore [26]. One of the major disadvantages besides low alignment with human preferences as shown by [27] for clinical text summarization is the need for human-generated reference summaries to evaluate against. This issue is aggravated in the healthcare domain, where physicians and other healthcare professionals have only limited resources for annotation. Recently, research in the domain of quality estimation (QE) has shifted toward developing and validating reference-free metrics, which do not require reference translations and instead assess the adequacy and fluency of the output directly [28]. Examples for reference-free metrics are YiSi-2 [29], Prism-src [30], SentSim [31], RefFreeEval [32], and MT-Ranker [33]. Another trend also explores the use of LLMs as automatic evaluators, also called “LLM-as-a-judge” [34] In the domain of MT evaluation, an LLM can be prompted to perform QE. While such approaches show strong correlation with human judgments, they are sensitive to prompt design and can exhibit systematic biases [35], [36].

In this paper, we assess the two QE metrics COMET-src and GEMBA-DA-noref. COMET-src is a reference-free variant of the widely used COMET metric. Both COMET and COMET-src have been shown to be among the best-performing MT evaluation metrics [37]. A specific COMET variant, CometKiwi-DA-XL [38], is used for the automated evaluation of the General Machine Translation Task at the 2024 Conference on Machine Translation (WMT) [39]. For our experiments, we use the largest available model variant of COMET-src, Unbabel/wmt23-cometkiwi-da-xxl, comprising 10.5 billion parameters and requiring a minimum of 44 GB of GPU memory [40].

GEMBA (GPT Estimation Metric Based Assessment) is an LLM-based evaluation approach that uses prompting to directly estimate translation quality. In its reference-free variant, GEMBA-DA-NoRef prompts the model to assign a direct assessment score between 0 and 100 without access to a reference translation and has shown strong system-level correlation with human judgments in prior evaluations on general-domain benchmarks [41]. Related approaches such as EAPrompt extend this idea by incorporating explicit error-aware prompting to obtain more fine-grained quality estimates, however at the cost of increased prompting complexity [42].

3 Results

3.1 Testing variants of the S-RAT pipeline

We tested four variants of the S-RAT pipeline (Base, Dict, Shot, and Curated). Each variant was evaluated on the same dataset with both gemma3 model variants (4b and 12b). Furthermore, the best-performing variant (Curated) was evaluated using two additional models (med42-v2:8b, deepseek-r1:8b). Additionally, we evaluated S-RAT-Base on gpt-oss:120b as large open-source model and Gemini 3 Pro as state-of-the-art commercial model, resulting in a total of twelve configurations.

S-RAT-Base uses the basic system prompt (Appendix G in Attachment 1) and includes all previously described S-RAT components, except the fallback translation list.

S-RAT-Dict extends the base pipeline by adding the fallback translation list (generation described in Appendix E in Attachment 1).

S-RAT-Shot applies a more complex system prompt, containing two manually curated translation examples for in-context learning.

S-RAT-Curated incorporates a manually curated dictionary of 173 concepts. This dictionary combines: (i) automatically translated and verified terms, (ii) entries from the fallback list (S-RAT-Dict), and (iii) frequent uni-, bi-, and trigrams from the dataset. In this variant, concept translations are added to the S-RAT-Shot prompt, but only for exact string matches.

For each configuration, the pipeline was applied to the full subset of 500 reports. Each report was translated based on two variants: once with keywords included in the prompt (-K, “S-RAT”) and once without (-NK, “standard”), yielding 1,000 translations per configuration.

We then computed the COMET-src score for every translation. ranging from 0 to 1 with higher values reflecting better translation quality. To evaluate whether our S-RAT approach differs from the standard approach, we applied a two-step evaluation strategy. For the primary analysis, we collapsed across model configurations: for each of the 500 clinical reports, we computed the mean COMET score across the eight models separately for the standard and S-RAT translation variants. These paired values were then compared using a Wilcoxon signed-rank test, providing an overall assessment of whether the S-RAT variant achieved higher translation quality.

For the secondary analysis, we compared the S-RAT and standard variants within each of the twelve model configurations using Wilcoxon signed-rank tests. Holm correction was applied to adjust for multiple comparisons. Effect sizes were reported as rank-based r. Together, these analyses tested both the overall benefit of the S-RAT variant and its consistency across different model architectures.

Result: Keyword-enhanced translation does not perform consistently better than standard prompting. In Table 1, we report the performance of each configuration. In the

primary analysis, the keyword variant did not significantly outperform the standard variant (Wilcoxon signed-rank test, W=62563.00, p=9.847e-01, r=0.001, n=500) averaged across all model configurations. In the secondary analysis, three variant comparisons reached significance after Holm correction. Baseline performance (gpt-oss:120b) decreased significantly using keyword-enhanced translation. Gemini 3 Pro did not outperform gemma3:12b, but its performance increased significantly using keyword-enhanced translation. The performance of the fine-tuned model med42-v2:8b showed overall low performance and significantly decreased using keyword-enhanced translation. Median differences favored the keyword-enhanced variant in seven out of twelve comparisons, see Appendix D (Attachment 1).

Table 1: Average COMET translation scores ($\bar{x}$) across model configurations, shown with (K) and without (NK) keyword conditioning (n=500). Best performance highlighted in bold.

Model configuration	$\bar{x}$ K	$\bar{x}$ NK
Baseline 1 (gpt-oss:120b)	0.69	**0.70**
Baseline 2 (Gemini 3 Pro)	0.67	0.66
S-RAT-Base (gemma3:4b)	0.62	0.62
S-RAT-Base (gemma3:12b)	0.66	0.67
S-RAT-Dict (gemma3:4b)	0.62	0.61
S-RAT-Dict (gemma3:12b)	0.67	0.66
S-RAT-Shot (gemma3:4b)	0.65	0.64
S-RAT-Shot (gemma3:12b)	0.68	0.68
S-RAT-Curated (gemma3:4b)	0.65	0.65
S-RAT-Curated (gemma3:12b)	0.68	0.68
S-RAT-Curated (med42-v2:8b)	0.60	0.62
S-RAT-Curated (deepseek-r1:8b)	0.64	0.63

3.2 Verifying SNOMED CT as feasible resource for concept translation

To verify whether SNOMED CT is a feasible resource for concept translation, we compute the observed translation coverage of our pipeline. We analyze the results of the S-RAT-Base configuration and obtain the total number of extracted concept instances (T). We then determine the number of instances that could not be translated (U) based on the list of untranslated concepts and their frequencies.

Our primary metric, the observed translation coverage, is calculated as:

$$\hat{\theta}_{obs} = 1 - \frac{U}{T}$$

We report this proportion with a 95% confidence interval and perform a one-sided exact binomial test to assess whether coverage exceeds the predefined 80% feasibility threshold.

To account for vocabulary breadth, we additionally compute distinct-type coverage, defined as the proportion of unique extracted concept types for which a German translation was available. We also present the distribution of untranslated concepts by frequency (Pareto analysis), highlighting whether some terms account for most untranslated instances.

This analysis measures only the proportion of extracted concepts covered by SNOMED CT. We do not evaluate the correctness of concept mappings or the clinical adequacy of translations due to the absence of human annotation. Results are therefore interpreted as coverage conditional on the output of the fixed extraction system. However, we perform a separate ablation study to assess recall and precision of the commercial entity recognition tool.

Result: The German SNOMED CT extension provides insufficient coverage for radiological use cases. A total of 6,305 concept instances were extracted from 500 radiology reports. Of these, 4,247 instances could not be translated into German, corresponding to an observed translation coverage of 32.64% (95% CI: 31.5–33.8%). This coverage falls well below the predefined feasibility threshold of 80%. Compared to S-RAT-Base, S-RAT-Dict improved translation coverage from 32.64% to 81.65%. S-RAT-Curated nearly quadrupled the average number of extracted concepts per report from 12.61 to 48.35. The most frequent untranslated terms included Chest (n=375), Consolidation (n=180), Male population group (n=155), and Abnormally opaque structure (n=150). A Pareto analysis showed that the top 20 untranslated terms accounted for the entire set of untranslated instances, with the ten most common terms already contributing over 70%. At the vocabulary level, 593 unique concept types lacked a German translation, resulting in a distinct-type coverage of 77.6%. This suggests that while most unique concept types were represented, frequent radiology terms remained systematically untranslated, disproportionately reducing instance-level coverage. Taken together, these results indicate that the German SNOMED CT edition, in its current form, does not provide sufficient coverage for translating radiology-related concepts from English to German.

3.3 Correlating automated QE metrics with human preferences

We base our human annotation methodology on the work of Chatzikoumi [43] One co-author, a radio-oncologist and therefore domain expert, performs a blinded assessment of the two translation variants (-K vs. -NK) based on the results of the S-RAT-Curated-4b configuration (n=199). For evaluation, the order of each triplet is randomly shuffled. Each triplet is comparatively rated for accuracy (= adequacy) using a bipolar scale ranging from –5 (strong preference for the -NK translation) to +5 (strong preference for the -K translation), based on existing work of Van Veen et al., who applied this approach

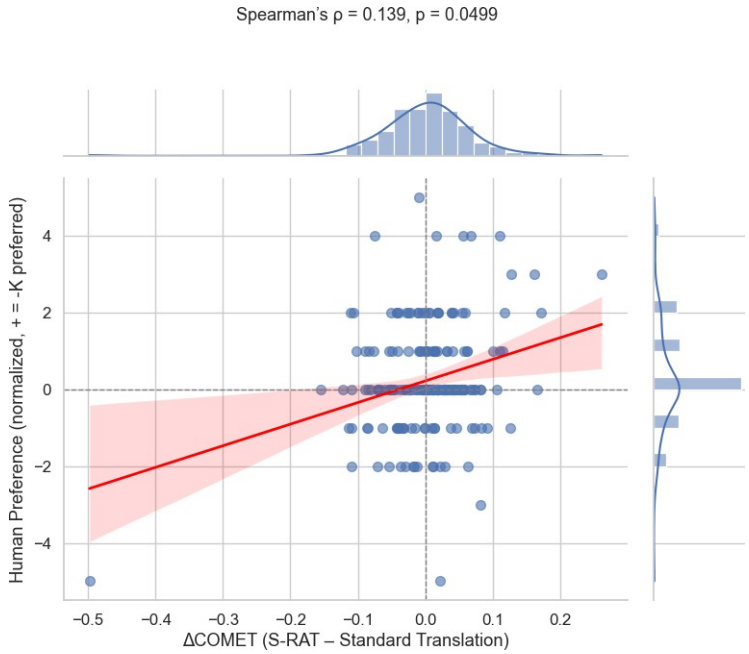

Figure 2: Relationship between human preference scores and COMET-src score differences (S-RAT minus standard prompting). Positive x-values indicate that COMET-src favoured S-RAT, whereas positive y-values indicate that the human expert preferred the S-RAT translation.

for comparison of human- and machine-generated summaries of clinical texts [27].

To compare these human preferences with each of the two automated metrics, we compute a difference score for each triplet by subtracting the QE score of the standard translation from that of the S-RAT translation, so that positive values indicate a higher QE score. Human ratings are treated as signed preference strengths. This allows direct comparison between human and metric scores: both are positive when S-RAT is preferred and negative when the standard translation is preferred.

To assess the degree to which COMET-src and GEMBA-DA-noref differences align with human preferences, we compute Spearman's rank correlation coefficient (Spearman's ρ) between the two vectors. In addition, we report the p-value associated with the correlation to determine whether the observed association is statistically significant.

Result: Automated metrics show limited alignment with human evaluation scores. The expert annotations (n=199) revealed a distribution where 47.2% of pairs were rated as equivalent, 31.2% favored S-RAT translations, and 21.6% favored standard translations. The mean rating of 0.231 (SD=1.399) indicates a slight overall preference for S-RAT. Spearman's rank correlation analysis revealed that COMET-src differences showed a weak but statistically significant correlation with human preferences (Spearman's ρ=0.139, p=0.0499, n=199), see Figure 2: Each point corresponds to one translated report pair, with the horizontal axis showing the COMET-src score difference (S-RAT minus standard prompting) and the vertical axis showing the corresponding human preference rating. In contrast, GEMBA-DA-noref differences showed no meaningful correlation (Spearman's ρ=0.007, p=0.917, n=198), despite its prior performance on general-domain MT benchmarks. While COMET-src captures a small fraction of the variance in expert adequacy judgments, GEMBA-DA-noref fails to align with human preferences at all. Given that both measures are intended to reflect translation quality, stronger correspondence would be expected if the metrics adequately modeled domain-specific adequacy. The weak alignment of COMET-src and the absence of correlation for GEMBA-DA-noref therefore indicate that these general-domain evaluation metrics do not fully reflect the clinical relevance and semantic precision valued by the human expert, highlighting a mismatch between general-domain evaluation metrics and domain-specific translation requirements.

3.4 Ablation study: Assessing the concept extraction tool

We manually evaluated a subset of 30 reports to assess the commercial entity recognition component. For each report, we annotated extracted concepts as true positives or false positives, and identified false negatives (missed concepts). From the 30 reports, the system extracted 419 concepts. Manual annotation showed 389 true positives and 30 false positives, yielding a precision of 92.8%. We identified 161 false negatives, resulting in a recall of 70.7% and an F1-score of 80.2%. The false positives comprised mostly incorrect normalization mappings (e.g., "LAT" → "ORC3 protein, human" instead of "lateral", "lead" → "Plumbum metallicum, homeopathic" instead of a medical device component, "MVR" → "Missing Value Reason" rather than "mitral valve replacement", "Heart" → "HEART PROBLEM").

The 161 false negatives comprised anatomical structures ("cardiac silhouette“, “hilar contours“), clinical findings ("pulmonary edema“, “atelectasis“, “pleural effusion“), medical devices (e.g., “dual chamber PPM“, “ETT placement“), and specific adjectives (e.g., “bronchovascular“, “costophrenic“). Many concepts that were missed in some reports were correctly identified in others, suggesting context-dependent recognition challenges rather than systematic failures, potentially aggravated by the rather low word count of radiology reports. See Appendix F (Attachment 1) for a detailed list of false positives and false negatives.

4 Conclusion

This study introduced Semantic Retrieval-Augmented Translation (S-RAT), a privacy-preserving approach that integrates structured clinical ontologies into machine translation workflows for clinical text. By leveraging UMLS and SNOMED CT concept mappings, S-RAT aims to improve terminology fidelity in radiology report translation without requiring cloud APIs or fine-tuning. We evaluated its feasibility, translation performance, and the reliability of automated evaluation metrics (COMET-src, GEMBA-DA-noref) against human judgment.

Overall, our findings were mixed. First, contrary to **H1**, the German edition of SNOMED CT did not provide sufficient concept coverage for radiological use cases, with only 32.64% of extracted concepts successfully translated. While most distinct concept types were represented, frequent high-value terms such as *Chest* and *Consolidation* were systematically untranslated, which strongly limited overall coverage. These findings highlight the importance of domain-specific curation of a dictionary (see S-RAT-Dict and S-RAT-Curated) before ontology-based translation can be deployed in production environments. The low coverage may partly reflect limitations of the German SNOMED CT edition rather than SNOMED CT itself. An edition with more complete target-language terminology could potentially yield different S-RAT results; however, we did not compare language editions directly. Future work should therefore assess whether terminology coverage is associated with translation performance.

Second, in relation to **H2**, retrieval-augmented prompting did not consistently outperform standard prompting. The lack of significant improvement might be driven by the prompting strategy itself. In several cases, manually curated or dictionary-based term injection helped preserve domain-specific phrasing but introduced occasional syntactic inconsistencies. Future work could explore adaptive prompt templates that dynamically adjust insertion granularity or confidence-weight concept injection based on retrieval certainty. Third, regarding **H3**, COMET-src correlated only weakly with expert adequacy ratings, while GEMBA-DA-noref showed no meaningful correlation. These findings indicate that general-domain reference-free metrics incompletely capture clinical translation quality. Although the weak yet statistically significant correlation suggests that COMET-src captures some degree of semantic alignment, it does not reflect the nuances of clinical correctness and factual precision that domain experts prioritize. This observation echoes similar findings in clinical summarization research, where general-domain metrics have shown poor alignment with human preferences for correctness and completeness [27]. Developing or fine-tuning MT quality estimation models on biomedical or radiological corpora could thus be a promising direction.

From a methodological perspective, the S-RAT framework demonstrates the feasibility of integrating structured knowledge into small, locally deployed LLMs. While the current implementation did not yield significant gains, it represents an important step toward data-sovereign medical translation, where hospitals can maintain control over sensitive data while still benefiting from LLM-based translation. The lightweight retrieval and prompting mechanism is model-agnostic and could be integrated into future hospital-internal translation workflows.

Future work could address several extensions, including the enrichment of SNOMED CT with missing radiological terminology via semi-automated LLM translation of uncovered concepts, incorporating human feedback to iteratively refine prompts, and exploring self-assessment mechanisms that enable the model to flag uncertain translations. Beyond radiology, the approach could be generalized to other domains (e.g., pathology or cardiology) where high-precision, ontology-linked translations are critical for patient safety.

In conclusion, while S-RAT did not yet outperform standard prompting in quantitative metrics, it provides a conceptual and technical foundation for ontology-augmented, privacy-compliant medical translation. The study highlights both the potential and the current infrastructural limitations of structured retrieval for LLM-driven clinical applications. Our findings call for stronger multilingual standardization efforts within SNOMED CT and more domain-specific evaluation frameworks for MT applied in healthcare settings.

Notes

Competing interests

The authors declare that they have no competing interests.

Authors' ORCIDs

- Daniel Reichenpfader: 0000-0002-8052-3359
- Fabio Dennstädt: 0000-0002-5374-8720

Author contributions

DR and FD contributed to the conception and execution of the study, as well as data collection and interpretation. DR was responsible for study design, data analysis, and

interpretation. DR drafted the manuscript, and FD critically revised it for important intellectual content. Both authors approved the final version of the manuscript and take responsibility for the scientific integrity of the work.

Attachments

Available from https://doi.org/10.3205/mibe000313

1. Attachment1_mibe000313.pdf (228 KB)
   Appendices A–G

References

1. Kong M, Fernandez A, Bains J, Milisavljevic A, Brooks KC, Shanmugam A, Avilez L, Li J, Honcharov V, Yang A, Khoong EC. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. BMJ Qual Saf. 2026 Feb 19;35(3):150-8. DOI: 10.1136/bmjqs-2024-018384
2. Team Gemma. Gemma 3 Technical Report [Preprint]. arXiv. 2025: arXiv:2503.19786 [cs]. DOI: 10.48550/arXiv.2503.19786
3. OpenAI, Agarwal S, Ahmad L, Ai J, Altman S, Applebaum A, et al. gpt-oss-120b & gpt-oss-20b Model Card [Preprint]. arXiv. 2025: arXiv:2508.10925 [cs]. DOI: 10.48550/arXiv.2508.10925
4. Apertus Team. Apertus: Democratizing open and compliant llms for global language environments. [last accessed 2026 Aug 20]. Available from: https://huggingface.co/swiss-ai/Apertus-70B-2509
5. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan;32(Database issue):D267-70. DOI: 10.1093/nar/gkh061
6. Meddeb A, Lüken S, Busch F, Adams L, Ugga L, Koltsakis E, Tzortzakakis A, Jelassi S, Dkhil I, Klontzas ME, Triantafyllou M, Kocak B, Yüzkan S, Zhang L, Hu B, Andreychenko A, Yurievich EA, Logunova T, Morakote W, Angkurawaranon S, Makowski MR, Wattjes MP, Cuocolo R, Bressem K. Large Language Model Ability to Translate CT and MRI Free-Text Radiology Reports Into Multiple Languages. Radiology. 2024 Dec;313(3):e241736. DOI: 10.1148/radiol.241736
7. Merx R, Suominen H, Cohn T, Vylomova E. OpenWHO: A Document-Level Parallel Corpus for Health Translation in Low-Resource Languages [Preprint]. arXiv. 2025: arXiv:2508.16048 [cs]. DOI: 10.48550/arXiv.2508.16048
8. García-Ferrero I, Agerri R, Salazar AA, Cabrio E, Iglesia Idl, Lavelli A, et al. Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain [Preprint]. arXiv. 2024: arXiv:2404.07613 [cs]. DOI: 10.48550/arXiv.2404.07613
9. Vo N, Nguyen DQ, Le DD, Piccardi M, Buntine W. Improving Vietnamese-English Medical Machine Translation [Preprint]. arXiv. 2024: arXiv:2403.19161 [cs]. DOI: 10.48550/arXiv.2403.19161
10. Rios M. Instruction-tuned Large Language Models for Machine Translation in the Medical Domain [Preprint]. arXiv. 2024: arXiv:2408.16440v1 [cs]. DOI: 10.48550/arXiv.2408.16440
11. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Preprint]. arXiv. 2021: arXiv:2005.11401 [cs]. DOI: 10.48550/arXiv.2005.11401
12. Wang J, Meng F, Zhang Y, Zhou J. Retrieval-Augmented Machine Translation with Unstructured Knowledge [Preprint]. arXiv. 2024: arXiv:2412.04342v1 [cs]. DOI: 10.48550/arXiv.2412.04342
13. Microsoft Azure. What is the Text Analytics for health in Azure AI Language? [last accessed 2026 Aug 20]. Available from: https://learn.microsoft.com/en-us/azure/ai-services/language-service/text-analytics-for-health/overview
14. NIH National Library of Medicine. UMLS API Home. [last accessed 2026 Aug 20]. Available from: https://documentation.uts.nlm.nih.gov/rest/home.html
15. SNOMED International. IHTSDO/snowstorm. Github; [last accessed 2026 Aug 20]. Available from: https://github.com/IHTSDO/snowstorm
16. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 2019 Dec;6(1):317. DOI: 10.1038/s41597-019-0322-0
17. Christophe C, Kanithi PK, Raha T, Khan S, Pimentel MA. Med42-v2: A Suite of Clinical LLMs [Preprint]. arXiv. 2024: arXiv:2408.06142 [cs]. DOI: 10.48550/arXiv.2408.06142
18. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Preprint]. arXiv. 2025: arXiv:2501.12948 [cs]. DOI: 10.48550/arXiv.2501.12948
19. Google DeepMind. Gemini 3 Pro Model Card. 2025 Nov [last updated 2026 May]. Available from: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf
20. Chen CL, Dong Y, Castillo-Zambrano C, Bencheqroun H, Barwise A, Hoffman A, Nalaie K, Qiu Y, Boulekbache O, Niven AS. A systematic multimodal assessment of AI machine translation tools for enhancing access to critical care education internationally. BMC Med Educ. 2025 Jul;25(1):1022. DOI: 10.1186/s12909-025-07452-9
21. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems 33. NeurIPS 2020. Curran Associates, Inc.; 2020 [last accessed 2026 Aug 20]. p. 1877-901. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
22. Reichenpfader D, Dennstädt F. RAT: Retrieval-augmented translation for medical texts (source code). Zenodo; 2026. DOI: 10.5281/zenodo.21995221
23. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2002. p. 311-8. DOI: 10.3115/1073083.1073135
24. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Association for Computational Linguistics; 2004 [last accessed 2026 Aug 20]. p. 74-81. Available from: https://aclanthology.org/W04-1013/
25. Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Goldstein J, Lavie A, Lin CY, Voss C, editors. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics; 2005 [last accessed 2026 Aug 20]. p. 65-72. Available from: https://aclanthology.org/W05-0909/
26. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT [Preprint]. arXiv. 2020: arXiv:1904.09675 [cs]. DOI: 10.48550/arXiv.1904.09675

27. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024 Apr;30(4):1134-42. DOI: 10.1038/s41591-024-02855-5

28. Zhao H, Liu Y, Tao S, Meng W, Chen Y, Geng X, et al. From Handcrafted Features to LLMs: A Brief Survey for Machine Translation Quality Estimation. In: 2024 International Joint Conference on Neural Networks (IJCNN); 2024 Jun 30 - Jul 05; Yokohama, Japan. IEEE; 2024. DOI: 10.1109/IJCNN60899.2024.10650457

29. Lo Ck. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In: Bojar O, Chatterjee R, Federmann C, Fishel M, Graham Y, Haddow B, et al, editors. Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). Association for Computational Linguistics; 2019. p. 507-13. DOI: 10.18653/v1/W19-5358

30. Thompson B, Post M. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2020. p. 90-121. DOI: 10.18653/v1/2020.emnlp-main.8

31. Song Y, Zhao J, Specia L. SentSim: Crosslingual Semantic Evaluation of Machine Translation. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tur D, Beltagy I, Bethard S, et al, editors. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics; 2021. p. 3143-56. DOI: 10.18653/v1/2021.naacl-main.252

32. Wu H, Han W, Di H, Chen Y, Xu J. A Holistic Approach to Reference-Free Evaluation of Machine Translation. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics; 2023. p. 623-36. DOI: 10.18653/v1/2023.acl-short.55

33. Moosa IM, Zhang R, Yin W. MT-Ranker: Reference-free machine translation evaluation by inter-system ranking [Preprint]. arXiv. 2024: arXiv:2401.17099 [cs]. DOI: 10.48550/arXiv.2401.17099

34. Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2023. p. 2511-22. DOI: 10.18653/v1/2023.emnlp-main.153

35. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23; 2023 Dec 10-16; New Orleans, Louisiana, USA. Red Hook, NY, USA: Curran Associates Inc.; 2023. p. 46595-623. DOI: 10.52202/075280-2020

36. Wang P, Li L, Chen L, Cai Z, Zhu D, Lin B, et al. Large Language Models are not Fair Evaluators. In: Ku LW, Martins A, Srikumar V, editors. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2024. p. 9440-50. DOI: 10.18653/v1/2024.acl-long.511

37. Kocmi T, Federmann C, Grundkiewicz R, Junczys-Dowmunt M, Matsushita H, Menezes A. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In: Barrault L, Bojar O, Bougares F, Chatterjee R, Costajussa MR, Federmann C, et al, editors. Proceedings of the Sixth Conference on Machine Translation. Association for Computational Linguistics; 2021 [last accessed 2026 Aug 20]. p. 478-94. Available from: https://aclanthology.org/2021.wmt-1.57

38. Rei R, Guerreiro NM, Pombal J, van Stigt D, Treviso M, Coheur L, et al. Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. In: Koehn P, Haddow B, Kocmi T, Monz C, editors. Proceedings of the Eighth Conference on Machine Translation. Association for Computational Linguistics; 2023. p. 841-8. DOI: 10.18653/v1/2023.wmt-1.73

39. Kocmi T, Avramidis E, Bawden R, Bojar O, Dvorkovich A, Federmann C, et al. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In: Haddow B, Kocmi T, Koehn P, Monz C, editors. Proceedings of the Ninth Conference on Machine Translation. Association for Computational Linguistics; 2024. p. 1-46. DOI: DOI: 10.18653/v1/2024.wmt-1.1

40. Unbabel. Unbabel/COMET. Github; [last accessed 2026 Aug 20]. Available from: https://github.com/Unbabel/COMET

41. Kocmi T, Federmann C. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In: Nurminen M, Brenner J, Koponen M, Latomaa S, Mikhailov M, Schierl F, et al, editors. Proceedings of the 24th Annual Conference of the European Association for Machine Translation. European Association for Machine Translation; 2023 [last accessed 2026 Aug 20]. p. 193-203. Available from: https://aclanthology.org/2023.eamt-1.19/

42. Lu Q, Qiu B, Ding L, Zhang K, Kocmi T, Tao D. Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models. In: Ku LW, Martins A, Srikumar V, editors. Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics; 2024. p. 8801-16. DOI: 10.18653/v1/2024.findings-acl.520

43. Chatzikoumi E. How to evaluate machine translation: A review of automated and human metrics. Natural Language Engineering. 2020 Mar;26(2):137-61. DOI: 10.1017/S1351324919000469

44. DiMascio C. cdimascio/py-readability-metrics. Github; [last accessed 2026 Aug 20]. Available from: https://github.com/cdimascio/py-readability-metrics

45. Chan YH. Biostatistics 104: correlational analysis. Singapore Medical Journal. 2003 Dec;44(12):614-9.

46. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018 Sep;18(3):91-3. DOI: 10.1016/j.tjem.2018.08.00

Corresponding author:

Prof. Daniel Reichenpfader
Berner Fachhochschule, Technik und Informatik, Forschung und Dienstleistung, Quellgasse 21, 2502 Biel/Bienne, Switzerland, Phone: +41 31 848 60 93
daniel.reichenpfader@bfh.ch

Please cite as

Reichenpfader D, Dennstädt F. Semantic retrieval-augmented translation for radiology reports using small language models: Algorithm development and evaluation. GMS Med Inform Biom Epidemiol. 2026;22:Doc15.
DOI: 10.3205/mibe000313, URN: urn:nbn:de:0183-mibe0003134

This article is freely available from

https://doi.org/10.3205/mibe000313

Published: *2026-09-22*