Supplementary material: Appendices A–G

A Licences and tools

For concept extraction, we used Microsoft's commercial product Azure Text Analytics for Health [13]. At the time of writing, the service could be deployed locally via a Docker container, ensuring that no data were transmitted to the cloud apart from usage quota information. With our experiments, we did not exceed the free tier. To retrieve the SNOMED CT identifiers of extracted concepts, we accessed the UMLS REST API, provided free of charge by the U.S. National Library of Medicine. Use of UMLS in software applications requires signing a license agreement [5]. Translations of identified concepts were obtained from a local Snowstorm installation, preloaded with the SNOMED CT Germany Edition, Version May 2025, V1.0. Access to the corresponding source files also required a license agreement.

B Data preprocessing

Python's random.sample() function was used to randomly select 500 indices from the total dataset, then created a subset using the dataset.select(indices) method. We applied random sampling with a fixed seed (random_seed = 42) to ensure reproducibility. The subset contains 495 unique patients. Reports average 644 characters and 89 words (range: 26–302 words). Figure S1 illustrates the distribution of report lengths. Most reports are relatively short, although a smaller number contain substantially more words. This variability reflects routine clinical reporting and demonstrates that the evaluation includes both concise and more complex reports. Readability analysis using the python package py-readability-metrics by DiMascio [44] shows an average Flesch Reading Ease score of 9.51 (range: 0-61.22), Flesch-Kincaid Grade Level of 3.42 (range: 0–20.01), and Automated Readability Index of 3.2 (range: 0–20.51). For the improved few-shot prompt, we manually selected two MIMIC-CXR reports and confirmed they were not part of the subset.

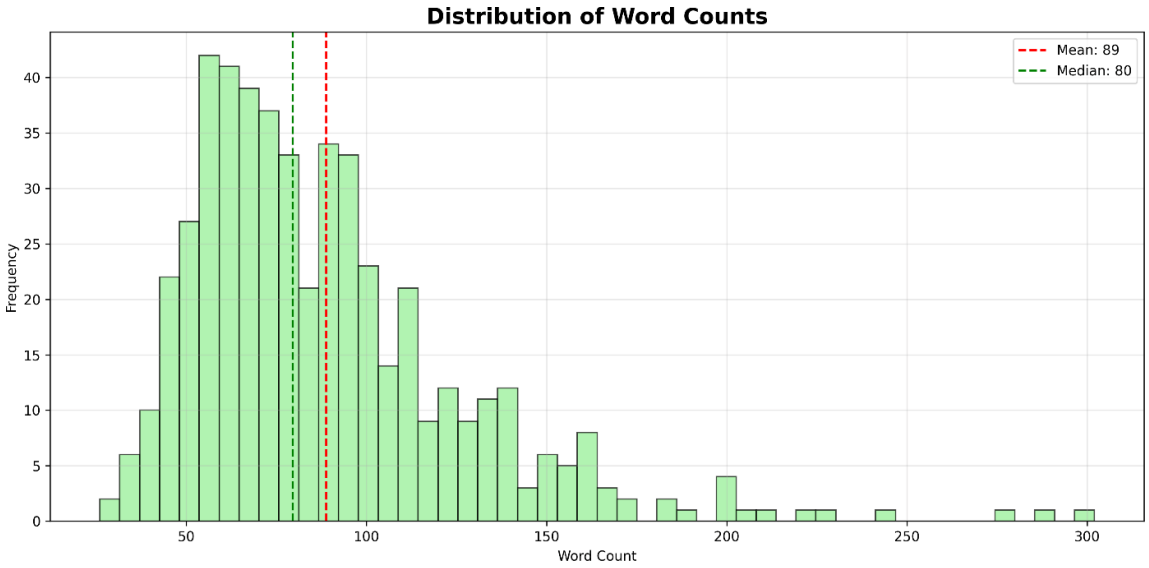

Figure S1: Distribution of word counts for the 500 randomly sampled MIMIC-CXR reports. Most reports contain fewer than 100 words, although a small number are substantially longer, illustrating the variability in report complexity used to evaluate S-RAT.

C Correlation computation

We interpret Spearman's ρ based on the work of Chan [45], compared with other interpretation scales, as shown by Akoglu [46].

D Results of secondary analysis

We provide a detailed overview of per-model comparison results in Table S1.

Attachment 1 to: 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

Table S1: Per-model Wilcoxon signed-rank tests comparing keyword-based (K) translations vs. standard (NK) translation COMET-src scores (n=500). r is a rank-based effect size; p_{Holm} is Holm-adjusted for multiple comparisons. Holm-adjusted p<0.05 shown in bold.

Model	*W*	p_{raw}	p_{Holm}	Median Δ (K–NK)	*r*
S-RAT-BASE-gemma3:4b	59468	0.329	1.000	0.0011	0.044
S-RAT-BASE-gemma3:12b	53880	0.0068	0.061	–0.0039	–0.121
S-RAT-DICT-gemma3:4b	61133	0.644	1.000	–0.0028	–0.021
S-RAT-DICT-gemma3:12b	60200	0.453	1.000	0.0014	0.034
S-RAT-SHOT-gemma3:4b	58859	0.244	1.000	0.0011	0.052
S-RAT-SHOT-gemma3:12b	62396	0.944	1.000	–0.0003	–0.003
S-RAT-CURATED-gemma3:4b	56774	0.070	0.452	0.0050	0.081
S-RAT-CURATED-gemma3:12b	56301	0.050	0.403	0.0029	0.087
S-RAT-CURATED-deepseek-r1:8b	56650	0.065	0.452	0.0017	0.083
S-RAT-CURATED-med42-v2:8b	**50559**	**0.00019**	**0.0021**	**–0.0083**	**–0.167**
S-RAT-BASE-gpt-oss-120b	**50934**	**0.00030**	**0.0030**	**–0.0053**	**–0.162**
S-RAT-BASE-gemini-3-pro	**40936**	**1.9e-11**	**2.3e-10**	**0.0089**	**0.300**

E Fallback translation list

The fallback translation list (used in S-RAT-Dict) is based on an analysis of failed translations from the S-RAT-Base configuration, and includes medical terms that could not be successfully translated through automated SNOMED CT lookups. We kept only the 100 most frequent terms and added manual translations for each extracted term. The dictionary is automatically integrated into the translation pipeline and serves as a fallback mechanism when SNOMED CT translation fails. During translation, if a concept extracted by Azure Text Analytics for Health cannot be translated via SNOMED CT (either directly or through UMLS lookup), the system queries this dictionary using exact term matching. Only entries marked as verified are used, ensuring translation quality and consistency across the dataset.

The comprehensive curated dictionary (used in S-RAT-Curated) extends this approach by defining 173 manually verified concepts that combine three sources: (i) entries from the fallback translation list described above, (ii) most frequent, successfully translated and verified terms, and (iii) frequent uni-, bi-, and trigrams extracted from the dataset through word frequency analysis, all based on the results of the S-RAT-Base configuration. The concept translations are added to the prompt only for exact string matches, bypassing the concept extraction and SNOMED CT lookup steps entirely.

F Concept extraction service: False positives and false negatives

Table S2 and Table S3 provide the complete lists of 15 unique false positives and false negatives (161 concepts, categorized by type) from the concept extraction evaluation.

*Table S2: De-duplicated **false positives** of 30 test cases identified by one co-author*

Extracted concept	Incorrect normalization
PA	Pulmonary artery pressure
LAT	ORC3 protein, human
generator	Generator Dosage Form
lead	Plumbum metallicum
homeopathic biliary stents	Insertion of biliary stent
carina	Structure of carina
nasogastric tube	Nasogastric tube procedures
line	Intravascular line
subclavian	Subclavicular approach
kinking	Pseudocoarctation
sob	Dyspnea
MVR	Missing Value Reason
AP	ampicillin
Single view	View
Heart	HEART PROBLEM

Attachment 1 to: 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

*Table S3: De-duplicated **false negatives** of 30 test cases, grouped by category, identified by one co-author*

Category	Missed concepts
Anatomical	apical, aorta, aorta knob, aortic knob, cardiac silhouette, hilar, hilar contours, left heart contour, lung, lungs, mediastinal, mediastinum, osseous, pleura, thoracolumbar junction, vertebral
Clinical	atelectasis, bilateral costophrenic angles, bronchovascular crowding, cardiopulmonary abnormality, compressive atelectasis, diffuse infiltrative ab-normality, effusion, effusions, infiltrate, peribronchial opacification, pleural effusion, pulmonary edema, pulmonary infiltrate, pulmonary vascular engorgement, pulmonary vascularity, resp. distress
Devices	dobhoff tube, dual chamber PPM, ET tube, Et tube tip, left subclavian line, pacer, posterior fixation hardware, radiopaque cylinder, right internal jugular line tip
Procedures	laparotomy, midline sternal wires, tricuspid annuloplasty, tricuspid valve sugery
Abbreviation	asbnl, AcS Ao, CVA, MVC, PNA, pna, POD 9, PPM, SVC, TVr
Other	calcification, calcified, elongated de-cending aorta, hernia, turtous, vascular congestion

G Prompts

We evaluated four combinations of system and user prompt variants: for each S-RAT setting (with keywords / without keywords), we applied both a base prompt and an improved prompt. In the examples below, prompt snippets only used in the 'with keywords' variant are highlighted in red.

Base system prompt:

You are a professional medical translator. Respond only with the translation.

Base user prompt:

Translate the following medical text from English to German. Important medical terms and their translations: <keyword_context> Text to translate: <text>
Provide a natural and fluent translation that correctly uses the provided medical term translations.
Maintain clinical accuracy and use standard medical terminology in the target language.

Improved prompts. For S-RAT-Shot, we refined both the system and user prompts by applying an in-context learning strategy. Specifically, two human-generated translations of MIMIC-CXR reports were included in the system prompt as examples. Since we cannot share these reports, we instead present below a publicly available sample report from the original MIMIC-CXR publication [16].

Improved system prompt:

You are a professional medical translator. Provide only the translation without any explanations. # Few-Shot Examples for Medical Translation with Keywords

The following examples demonstrate proper medical translation from English to German with medical term enhancement:

Example 1:

Translate the following medical text from English to German.

Medical term translations:

- CHEST (English) → Thorax (German) - PA (English) → p.a. (German) - LAT (English) → lateral (German) - woman (English) → Frau (German) - pleural effusion (English) → Pleuraerguss (German) - pleural (English) → Pleural (German) - lateral (English) → Lateral (German) - effusion (English) → Erguss (German) - opacities (English) → Opazitäten (German) pneumothorax (English) → Pneumothorax (German) - degenerative changes (English) → Degenerative Veränderungen (German) - thoracic spine (English) → Brustwirbelsäule (German)

Attachment 1 to: 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

Text to translate:

EXAMINATION: CHEST (PA AND LAT)

INDICATION: _____year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

COMPARISON: _____

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine.

IMPRESSION:

Large left pleural effusion

Translated text:

UNTERSUCHUNG: Thorax (p.a. und lateral)

INDIKATION: _____-jährige Frau mit ?Pleuraerguss // ?Pleuraerguss

TECHNIK: Thorax p.a. und lateral

VERGLEICH: _____

BEFUND:

Die Herzgröße kann nicht beurteilt werden. Ein großer linker Pleuraerguss ist neu. Ein kleiner rechter Erguss ist neu. Die oberen Lungen sind klar. Rechte-untere-Lappen-Opazitäten sind im vorherigen CT besser zu sehen. Es gibt keinen Pneumothorax. Es gibt milde degenerative Veränderungen in der Brustwirbelsäule.

BEURTEILUNG:

Großer linker Pleuraerguss

—

Example 2:

<left out>

—

Translation Guidelines

Based on these examples, follow these principles:
1. Maintain clinical accuracy and use standard medical terminology
2. Preserve the formal, professional tone of medical reports
3. Keep anatomical and diagnostic terms precise and consistent
4. Maintain the structured format of medical reports
5. Use the provided medical term translations when available
6. Ensure all medical codes and technical terms are properly translated

Improved user prompt:

Translate the following medical text from English to German.
Important medical terms and their translations: <keyword_context>
Text to translate: <text>
Translated text:

GEMBA Prompt. For the evaluation, we used the exact same prompt as specified by Kocmi and Federmann [41]:

Score the following translation from English to German on a continuous scale from 0 to 100, where score of zero means “no meaning preserved“ and score of one hundred means “perfect meaning and grammar“.
English source: <text>
German translation: <translation>
Score:

Attachment 1 to: 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