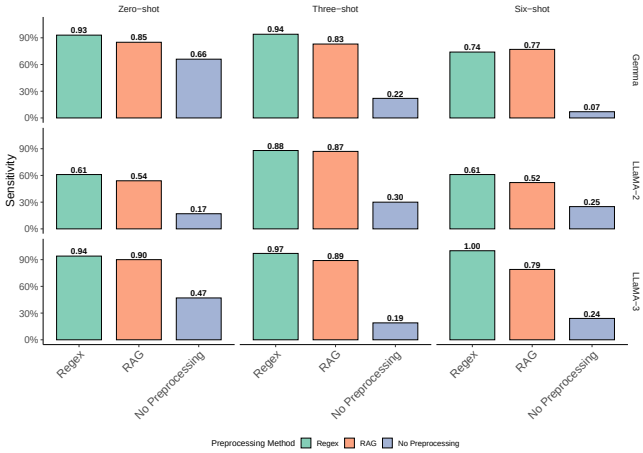
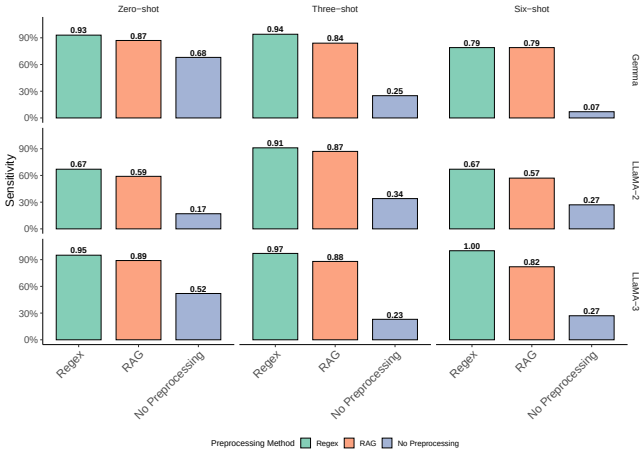
Supplementary Information

Contents

1 Additional Figures

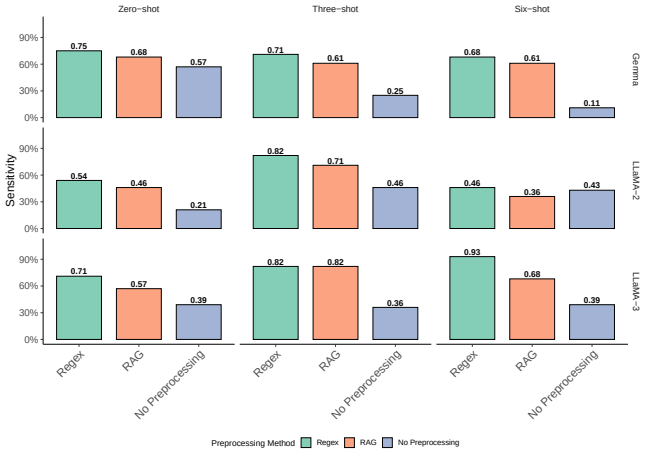


(a) Model sensitivity for metastasis detection within a 30-day time window on the HNC subset.

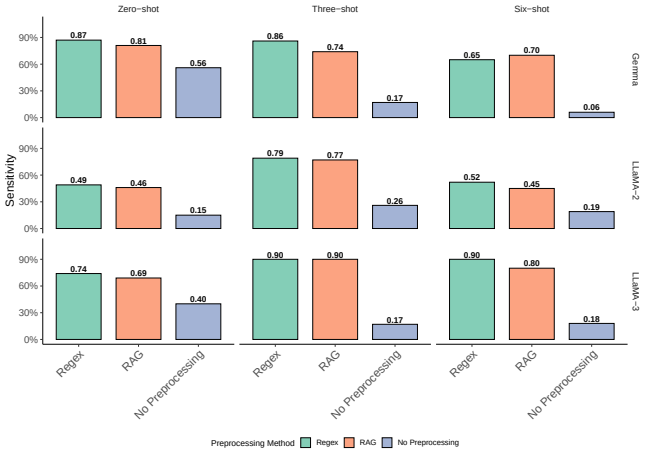


(b) Model sensitivity for metastasis detection within a 40-day time window on the HNC subset.

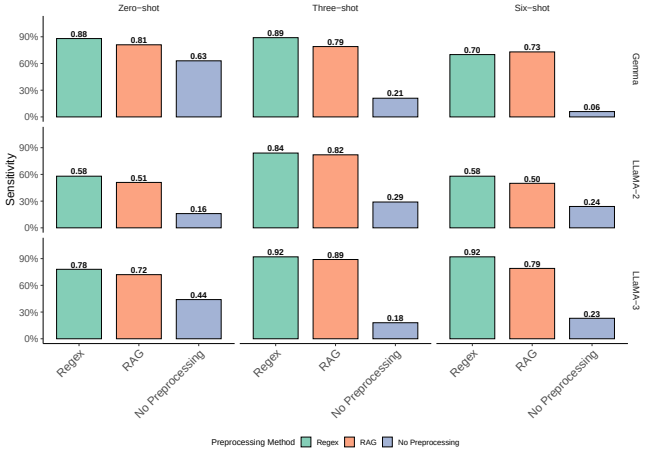
Supplementary Fig. 1: Model sensitivity for metastasis detection on the private HNC dataset subset, evaluated under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods across 30-day and 40-day time windows.



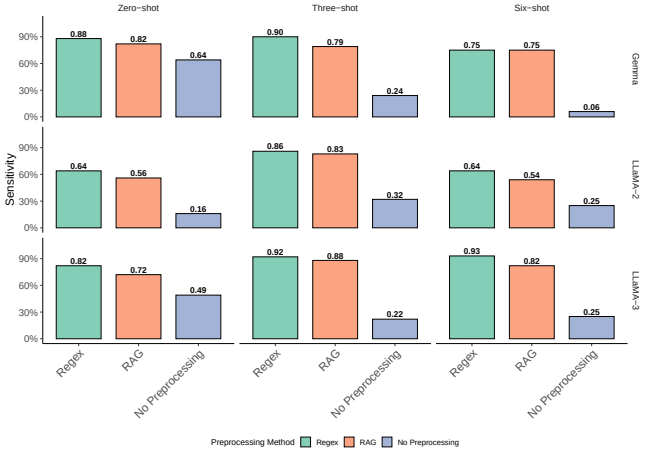
(a) System sensitivity for metastasis detection at the subject level on the HNC subset.



(b) System sensitivity for metastasis detection within a 20-day time window on the HNC subset.

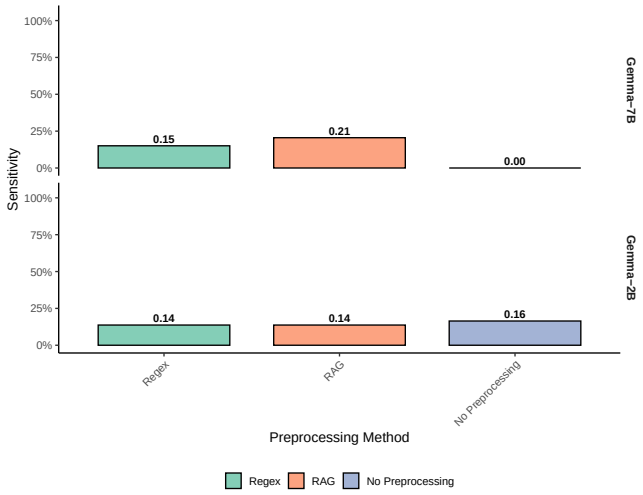


(c) System sensitivity for metastasis detection within a 30-day time window on the HNC subset.

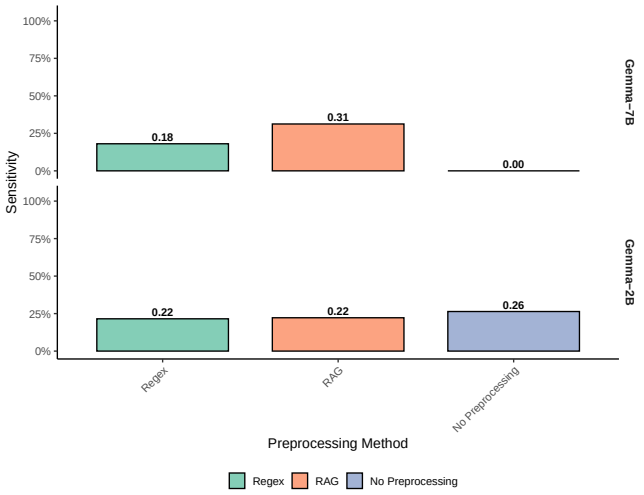


(d) System sensitivity for metastasis detection within a 40-day time window on the HNC subset.

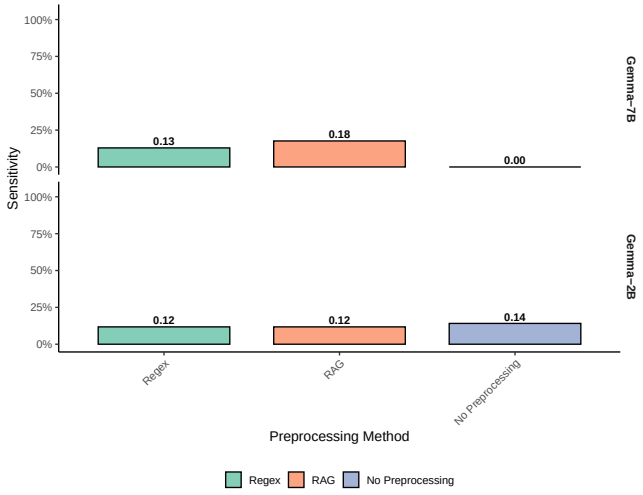
Supplementary Fig. 2: System sensitivity for metastasis detection on the private HNC dataset subset at the subject level and visit level across 20, 30, and 40-day time windows, evaluated under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods.



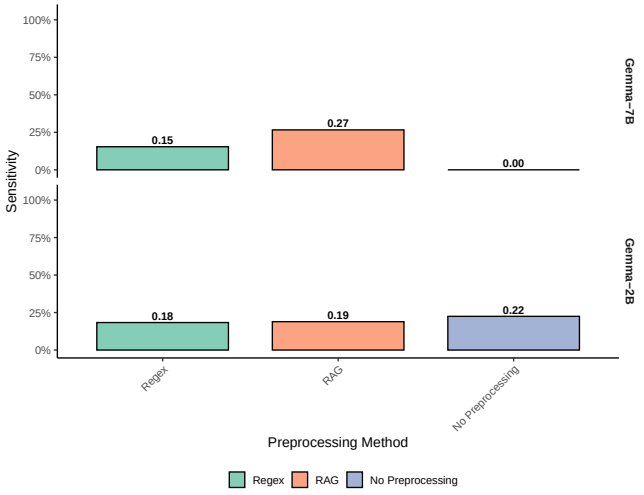
(a) Model sensitivity for metastasis detection at the subject level using fine-tuned Gemma models.



(b) Model sensitivity for metastasis detection at the hospital admission level using fine-tuned Gemma models.

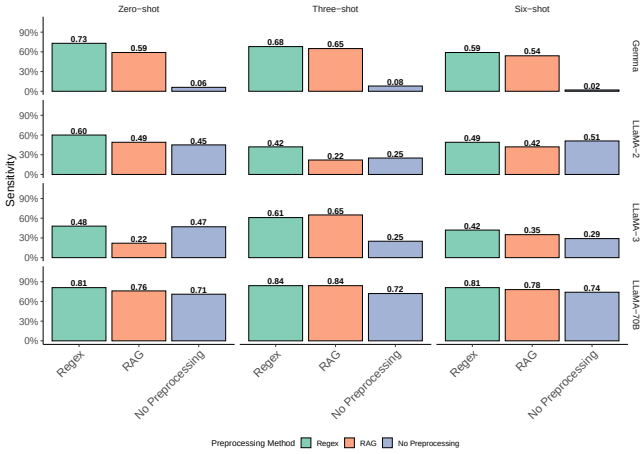


(c) System sensitivity for metastasis detection at the subject level using fine-tuned Gemma models.

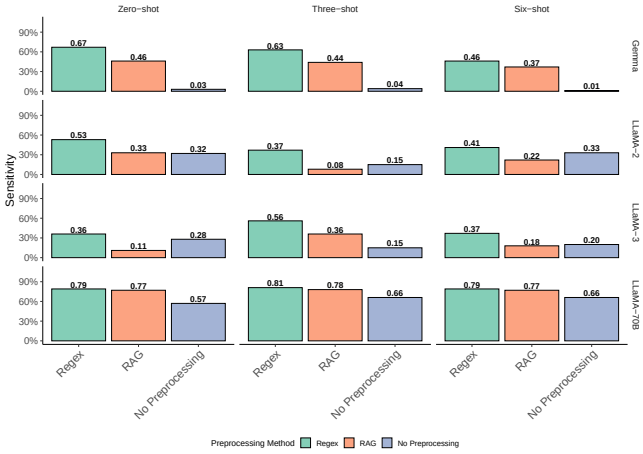


(d) System sensitivity for metastasis detection at the hospital admission level using fine-tuned Gemma models.

Supplementary Fig. 3: Model and system sensitivity for metastasis detection on the MIMIC-IV subset using fine-tuned `Gemma-2B-it` and `Gemma-7B-it` models, reported at both the subject level and the hospital admission level.

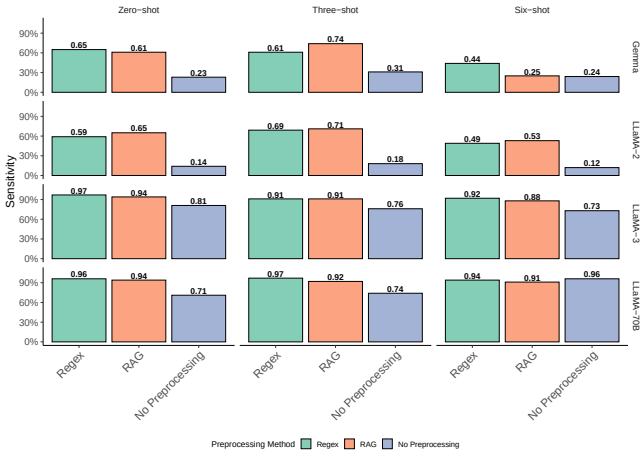


(a) System sensitivity for metastasis detection at the subject level on the MIMIC-IV subset.

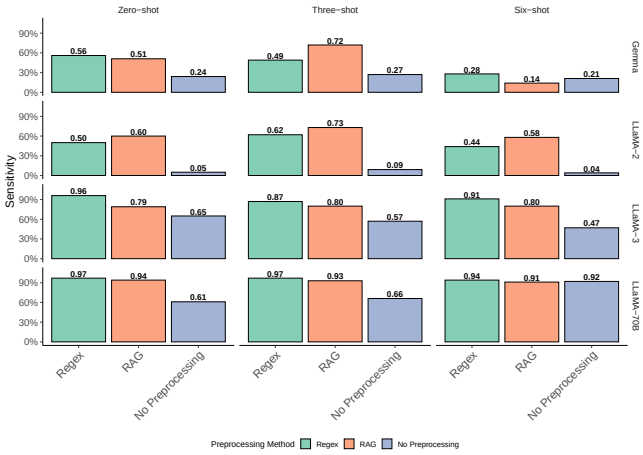


(b) System sensitivity for metastasis detection at the hospital admission level on the MIMIC-IV subset.

Supplementary Fig. 4: System sensitivity for metastasis detection on the MIMIC-IV subset at the subject level and the hospital admission level, evaluated across four LLMs under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods.

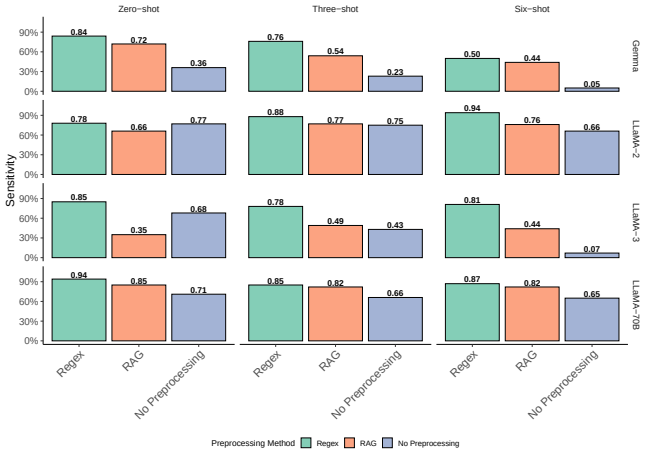


(a) Model sensitivity for insulin use detection at the subject level on the MIMIC-IV subset.

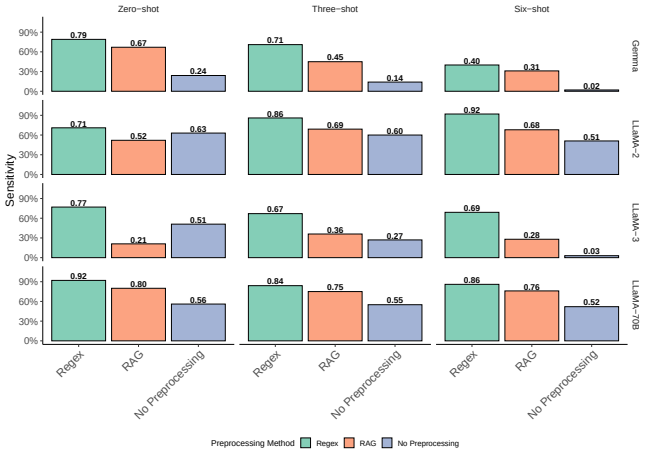


(b) Model sensitivity for insulin use detection at the hospital admission level on the MIMIC-IV subset.

Supplementary Fig. 5: Model sensitivity for insulin use detection on the MIMIC-IV subset at the subject level and the hospital admission level, evaluated across four LLMs under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods.

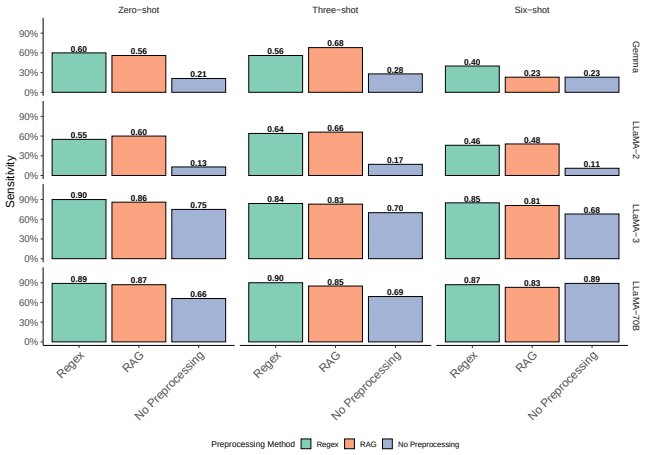


(a) Model sensitivity for hypertension detection at the subject level on the MIMIC-IV subset.

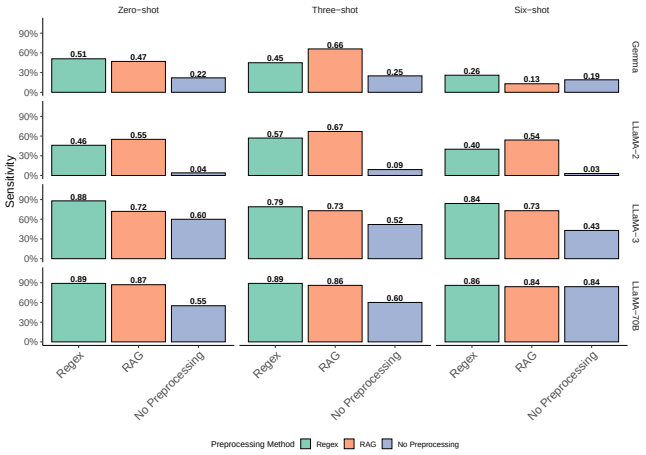


(b) Model sensitivity for hypertension detection at the hospital admission level on the MIMIC-IV subset.

Supplementary Fig. 6: Model sensitivity for hypertension detection on the MIMIC-IV subset at the subject level and the hospital admission level, evaluated across four LLMs under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods.

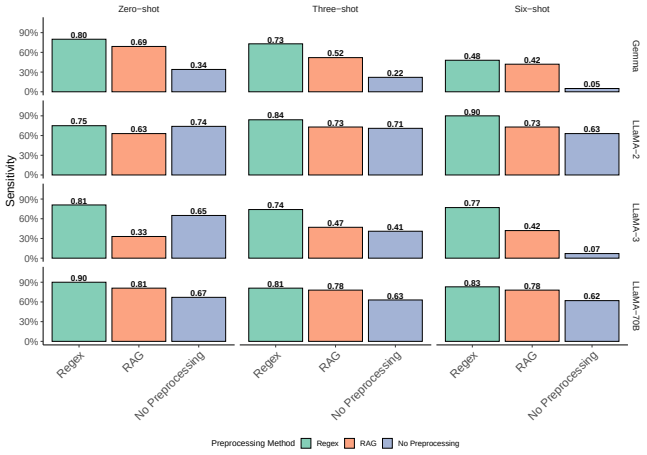


(a) System sensitivity for insulin use detection at the subject level on the MIMIC-IV subset.

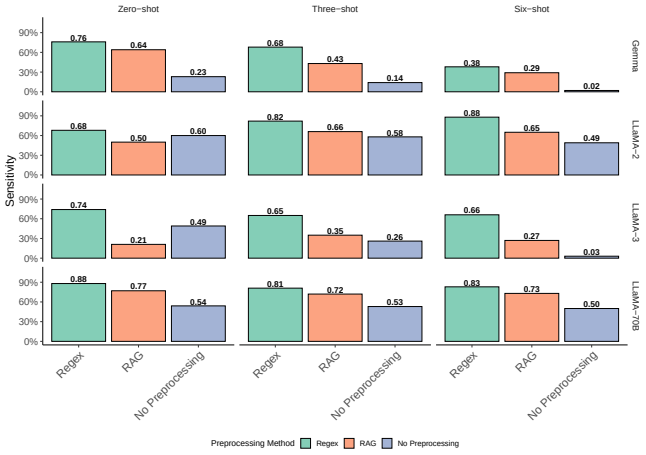


(b) System sensitivity for insulin use detection at the hospital admission level on the MIMIC-IV subset.

Supplementary Fig. 7: System sensitivity for insulin use detection on the MIMIC-IV subset at the subject level and the hospital admission level, evaluated across four LLMs under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods.



(a) System sensitivity for hypertension detection at the subject level on the MIMIC-IV subset.



(b) System sensitivity for hypertension detection at the hospital admission level on the MIMIC-IV subset.

Supplementary Fig. 8: System sensitivity for hypertension detection on the MIMIC-IV subset at the subject level and the hospital admission level, evaluated across four LLMs under zero-shot, three-shot, and six-shot settings with regex, RAG, and non-preprocessed methods.

The following sections provide supplementary details referenced in the main text, including LLM specifications and RAG implementation details (Supplementary Note 2), the keyword lists used for regex-based preprocessing (Supplementary Note 3), the ICD code definitions adopted as ground-truth labels (Supplementary Note 4), and the prompt templates used for LLM inference (Supplementary Note 5).

2 LLM Specifications and RAG Implementation Details

2.1 Gemma-7B-it

The primary small LLM used in PrecLLM was `Gemma-7B-it` **Version 1** (released February 2024 by Google DeepMind), the instruction-tuned variant with 7 billion parameters. The model weights were obtained via Kaggle Hub (`google/gemma/pyTorch/7b-it`) and deployed locally on NVIDIA GPUs using the official `gemma_pytorch` library in bfloat16 precision. The instruction-tuned variant was chosen over the base model because it was optimized for following structured prompts and producing constrained outputs, which was essential for the three-way classification task. Generation parameters were set as follows: temperature = 0.1, top-p = 1.0, top-k = 100, and maximum output length = 3 tokens. The low temperature was chosen to produce near-deterministic outputs suitable for classification.

2.2 LLaMA-Family Models

Three additional LLaMA-family models were evaluated to assess the generalizability of PrecLLM across different model architectures and scales. All three were loaded via the Hugging Face `transformers` library (`AutoModelForCausalLM`) with quantization provided by `bitsandbytes`, and shared the same core generation settings: `do_sample=True`, temperature = 0.1, and `max_new_tokens` = 4.

- **LLaMA-2-7B-Chat-Med** (`lianggq/llama-2-7b-chat-med`): A medical-domain fine-tune of LLaMA 2 with 7B parameters. Loaded in 8-bit quantization on a single GPU.
- **Bio-Medical-LLaMA-3-8B** (`ContactDoctor/Bio-Medical-Llama-3-8B`): A biomedical fine-tune of LLaMA 3 with 8B parameters. Loaded in 8-bit quantization on a single GPU.
- **Meta-Llama-3-70B-Instruct** (`meta-llama/Meta-Llama-3-70B-Instruct`): The instruction-tuned LLaMA 3 model with 70B parameters, included to evaluate PrecLLM at a larger scale. Loaded in 8-bit quantization with automatic multi-GPU device mapping.

2.3 RAG-Based Preprocessing Setup

As described in the main text, the RAG-based preprocessing served as an alternative to regex when domain knowledge for keyword curation was limited. The semantic retrieval component was implemented using the `sentence-transformers` library (Hugging Face) with the `all-MiniLM-L6-v2` model, a lightweight sentence embedding model based on Microsoft's MiniLM architecture that produces 384-dimensional dense vectors. Nearest-neighbor retrieval was performed using FAISS (Facebook AI Similarity Search) with an `IndexFlatL2` index, computing L2 (Euclidean) distances between sentence embeddings and the target query embedding.

The key hyperparameters for RAG-based retrieval were:

- **Top-k sentences**: Number of candidate sentences retrieved per note. Two settings were evaluated: $k = 2$ (denoted RAG2) and $k = 6$ (denoted RAG6).
- **Distance threshold**: 1.5 (L2 distance). A sentence was retained only if its distance to the query embedding was below this threshold, even when ranked in the top k.

For each clinical note, sentences were first segmented using regex-based boundary detection (splitting on double newlines or sentence-ending punctuation), with short fragments (< 5 characters) merged into adjacent sentences. Each sentence was then encoded into a 384-dimensional embedding, and FAISS retrieved the top-k most similar sentences to the query. Only sentences passing both the top-k and distance-threshold criteria were forwarded to the LLM for classification.

3 Keyword Lists for Regex-Based Preprocessing

As described in the main text, the regex-based preprocessing step relies on a curated set of Semantically Similar Terms (SSTs) for each target clinical variable. These keyword lists were constructed through a combination of clinical domain expertise and generative AI assistance (e.g., GPT-4), following three guiding principles: (1) inclusion of canonical medical terms and their morphological variants (e.g., "metastasis", "metastatic", "metastasized"); (2) coverage of common clinical abbreviations and shorthand notations (e.g., "HTN" for hypertension, "IDDM" for insulin-dependent diabetes mellitus); and (3) incorporation of semantically related concepts that frequently co-occur with the target variable in clinical narratives (e.g., "hematogenous spread" for metastasis). The keywords were organized into a single regex search command using two complementary matching strategies: flexible patterns for structural variations and comprehensive synonym lists.

3.1 Metastasis Keywords

The following keywords capture the broad spectrum of terminology used to describe metastatic disease in clinical notes, ranging from primary diagnostic terms to specific pathological mechanisms:

- **Metastasis**, **Metastatic**, **Metastasize**, **Metastases**, **Metastasized**: Primary terms and morphological variants describing the spread of cancer from the original site to other parts of the body.
- **Dissemination**, **Distant spread**: General terms indicating cancer spread beyond local or regional boundaries.
- **Metachronous**, **Hematogenous spread**, **Lymphatic spread**: Terms specifying the temporal pattern or anatomical pathway of metastatic dissemination.
- **Micrometastases**, **Infiltration**: Terms referring to microscopic tumor deposits and tissue penetration, respectively.
- **Tumor spread**, **Extra-nodal extension**: Terms describing progressive tumor expansion and extension beyond the lymph node capsule.

3.2 Hypertension Keywords

Hypertension is documented in clinical notes using a relatively small but highly standardized set of terms and abbreviations:

- **Hypertension**, **HTN**: Primary terms denoting elevated blood pressure as a chronic condition.
- **BP elevated**, **Blood pressure elevated**: Descriptive phrases indicating acute or chronic elevation in blood pressure measurements.
- **Hypertensive**: Adjective form used to describe conditions, crises, or states related to high blood pressure (e.g., hypertensive emergency, hypertensive disorder).

3.3 Insulin Use Keywords

Keywords for insulin use target both the therapeutic intervention itself and the diagnostic labels that imply insulin dependence:

- **Insulin**, **Insulin therapy**: Primary terms indicating use of exogenous insulin for glycemic control.
- **Insulin-dependent**, **IDDM**: Terms describing insulin-dependent diabetes mellitus (Type 1 diabetes or insulin-requiring Type 2 diabetes).
- **Insulin regimen**, **Insulin dosing**: Phrases related to insulin treatment protocols and dosage adjustments.

4 ICD Code Definitions for Ground-Truth Labels

As noted in the main text, the International Classification of Diseases (ICD) codes served as the practical reference standard (proxy for ground truth) for evaluating PrecLLM's annotation accuracy. Both ICD-9-CM and ICD-10-CM code sets were required because the datasets span the United States coding transition: the private HNC dataset covers 2014–2022 and MIMIC-IV covers 2008–2019, whereas the mandatory switch from ICD-9-CM to ICD-10-CM occurred on October 1, 2015. For each target variable, codes were selected using prefix-based matching on well-established clinical categories, with selection criteria reviewed by domain clinicians to ensure clinical consistency. Supplementary Tables 1–3 list the complete set of codes used.

Supplementary Table 1: ICD codes used to identify metastatic conditions.

Version	Code	Description
ICD-9	197	Secondary malignant neoplasm of respiratory and digestive systems
ICD-9	198	Secondary malignant neoplasm of other specified sites
ICD-9	199	Malignant neoplasm without specification of site
ICD-10	C78	Secondary malignant neoplasm of respiratory and digestive organs
ICD-10	C79	Secondary malignant neoplasm of other and unspecified sites
ICD-10	C80	Malignant neoplasm without specification of site

Supplementary Table 2: ICD codes used to identify insulin use.

Version	Code	Description
ICD-9	V58.67	Long-term (current) use of insulin
ICD-10	Z79.4	Long-term (current) use of insulin

Supplementary Table 3: ICD codes used to identify hypertension.

Version	Code	Description
ICD-9	401	Essential hypertension
ICD-9	402	Hypertensive heart disease
ICD-9	403	Hypertensive chronic kidney disease
ICD-9	404	Hypertensive heart and chronic kidney disease
ICD-9	405	Secondary hypertension
ICD-10	I10	Essential (primary) hypertension
ICD-10	I11	Hypertensive heart disease
ICD-10	I12	Hypertensive chronic kidney disease
ICD-10	I13	Hypertensive heart and chronic kidney disease
ICD-10	I15	Secondary hypertension

5 Prompt Templates for LLM Inference

This section presents the prompt templates used for the LLM classification step (Step 2 of the PrecLLM pipeline). Each prompt was designed following three principles: (1) framing the task as a structured three-way classification ("Yes", "No", "Unknown") to reduce output variability and facilitate downstream majority voting; (2) restricting the model to output only a single numerical code, thereby minimizing hallucinated explanations; and (3) providing explicit instructions to default to "Unknown" when the evidence is ambiguous, which helps preserve the integrity of the majority-vote aggregation. The templates below use metastasis as the illustrative target variable; for other variables (insulin use, hypertension), only the variable name and the in-context examples are substituted while the template structure remains identical. In the few-shot setting, three or six representative examples were manually selected from the training pool to cover affirmative, negative, and ambiguous cases, ensuring the model was exposed to the full range of expected outputs.

5.1 Zero-Shot Prompt Template

Zero-shot learning task

Task: Classify the presence of metastasis from patient clinical notes. Respond only with the following numerical codes based on the notes provided:
- (1) Yes: The notes explicitly confirm the patient has metastasis. - (2) No: The notes explicitly confirm the patient does not have metastasis. - (3) Unknown: The notes do not contain sufficient information to determine the patient's metastasis status.
Instructions: 1. Do not provide explanations or reasons for your classification. 2. Use only the information in the notes for your classification. 3. If the notes are ambiguous or lack details regarding metastasis, choose "(3) Unknown". 4. Only provide a single numerical code as the response.
Examples:
- Example 1: "The CT scan shows multiple nodules in the liver consistent with metastasis." – (1)
- Example 2: "Patient has a history of cancer but no evidence of metastatic disease on recent imaging." – (2)
- Example 3: "Patient presents for cancer follow-up." – (3)
Analyze the following clinical notes to determine if the patient has metastasis:
{INSERT CURRENT CLINICAL NOTES HERE}

5.2 Few-Shot Prompt Template

In the few-shot setting, representative annotated examples are inserted before the query segment. The number of examples ($n = 3$ or $n = 6$) is varied to evaluate the trade-off between additional context and input length constraints of smaller LLMs.

Few-shot learning task

Task: Classify the presence of metastasis from patient clinical notes. Respond only with the following numerical codes based on the notes provided:
- (1) Yes: The notes explicitly confirm the patient has metastasis. - (2) No: The notes explicitly confirm the patient does not have metastasis. - (3) Unknown: The notes do not contain sufficient information to determine the patient's metastasis status.
Instructions: 1. Do not provide explanations or reasons for your classification. 2. Use only the information in the notes for your classification. 3. If the notes are ambiguous or lack details regarding metastasis, choose "(3) Unknown". 4. Only provide a single numerical code as the response.
Examples:
{INSERT n ANNOTATED EXAMPLES HERE}
Analyze the following clinical notes to determine if the patient has metastasis:
{INSERT CURRENT CLINICAL NOTES HERE}