Modeling Clinical Uncertainty in Radiology Reports: from Explicit Uncertainty Markers to Implicit Reasoning Pathways

Paloma Rabaey[1*], Jong Hak Moon[2*], Jung-Oh Lee[3], Min Gwan Kim[4], Hangyul Yoon[2], Thomas Demeester[1], Edward Choi[2]

[1]Ghent University – imec, [2]KAIST, [3]Mount Sinai Hospital, [4]Seoul National University Hospital

paloma.rabaey@ugent.be, jhak.moon@kaist.ac.kr

[*]Joint first authors.

Abstract

Radiology reports are invaluable for clinical decision-making and hold great potential for automated analysis when structured into machine-readable formats. These reports often contain uncertainty, which we categorize into two distinct types: (i) Explicit uncertainty reflects doubt about the presence or absence of findings, conveyed through *hedging* phrases. These vary in meaning depending on the context, making rule-based systems insufficient to quantify the level of uncertainty for specific findings; (ii) Implicit uncertainty arises when radiologists omit parts of their reasoning, recording only key findings or diagnoses. Here, it is often unclear whether omitted findings are truly absent or simply unmentioned for brevity. We address these challenges with a two-part framework. We quantify *explicit uncertainty* by creating an expert-validated, LLM-based *reference ranking* of common hedging phrases, and mapping each finding to a probability value based on this reference. In addition, we model *implicit uncertainty* through an expansion framework that systematically adds characteristic sub-findings derived from expert-defined diagnostic pathways for 14 common diagnoses. Using these methods, we release LUNGUAGE^{++}, an expanded, uncertainty-aware version of the LUNGUAGE benchmark of fine-grained structured radiology reports. This enriched resource enables uncertainty-aware image classification, faithful diagnostic reasoning, and new investigations into the clinical impact of diagnostic uncertainty.

Keywords: Radiology report, Chest X-ray, Clinical uncertainty, Diagnostic reasoning, Large language model

1. Motivation and Related Work

Radiology reports play a central role in clinical decision-making, serving as the primary medium through which radiologists communicate their interpretations and diagnostic impressions to referring physicians. These reports influence downstream diagnostic reasoning, and often determine treatment trajectories. As AI models have been increasingly used for both automated radiology report interpretation and generation, structuring frameworks have been developed to convert free-text reports into machine-readable formats that can be used for training and evaluation (Moon et al., 2026; Wu et al., 2021; Jain et al., 2021; Khanna et al., 2023). A key challenge that needs to be addressed is that **radiology reports inherently contain uncertainty**. This uncertainty arises in two distinct forms: explicit uncertainty, expressed directly through the language radiologists use to qualify their findings or diagnoses (Bruno et al., 2017; Hobby et al., 2000), and implicit uncertainty, which emerges from the selective and often incomplete nature of what is recorded in the report (Turner et al., 2021). Figure 1 shows how these two forms of uncertainty appear in chest X-ray (CXR) reports, which are especially prone to contain uncertainty (Callen et al., 2020; Irvin et al., 2019).

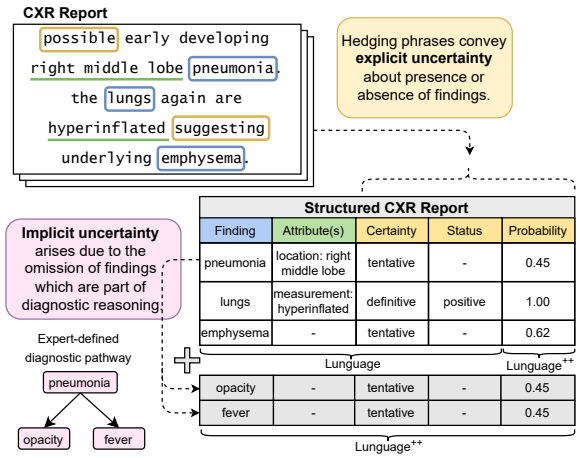


Figure 1: Two types of uncertainty in radiology reports that we address during structuring, expanding the LUNGUAGE dataset of structured CXR reports to form LUNGUAGE^{++}. Explicit uncertainty is conveyed by hedging phrases that indicate tentative findings, whose (un)certainty we quantify with probabilities. Implicit uncertainty stems from findings that are not explicitly mentioned; we mitigate this by applying expert-defined diagnostic pathways to expand stated diagnoses with their characteristic sub-findings.

Explicit uncertainty arises when radiologists convey doubt **about the presence or absence of a finding or diagnosis**, typically through *hedging* phrases such as "probably", "possible", "suggesting", or "may represent", among many others. Prior work examining these expressions has highlighted their prevalence, revealing the mention of uncertain diagnoses in a high proportion of CXR reports (Irvin et al., 2019; Moon et al., 2026). The use of hedging in radiology reports is deliberate and meaningful: it enables radiologists to communicate diagnostic

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11852–11873
11–16 May 2026.

uncertainty that is crucial for appropriate clinical interpretation and decision-making (Reiner, 2018; Bruno et al., 2017). Automated systems should similarly be designed to recognize and account for this uncertainty, just as human readers of radiology reports are trained to do.

Previous work has extracted uncertainty using rule-based systems (Irvin et al., 2019; Johnson et al., 2019a; Zhang et al., 2023) in ternary classification settings (positive / uncertain / negative). These approaches are limited in scope, using predefined vocabularies of hedging phrases that trigger an uncertain label when detected in the report. In this work, we go beyond discrete labeling by **quantifying uncertainty as a continuous probability between 0 and 1**, taking into account the specific hedging phrases and their sentence-level context (see Figure 1). Earlier attempts to assign probabilistic meanings to hedging expressions have typically relied on human judgments, asking experts to rate phrases or position them along a scale of certainty (Hobby et al., 2000; Shinagare et al., 2019). However, these approaches proved unreliable, as the interpretation of hedging varies widely across radiologists. To address this, we introduce an automated approach that estimates the probability of a finding by leveraging large language models to perform pairwise comparisons of uncertainty expressions, constructing a relative ranking that is then mapped to a continuous probability.

Beyond overt language, reports also embody a subtler form of **implicit uncertainty** that has received little systematic attention. Radiologists frequently **omit portions of their diagnostic reasoning, documenting only key findings** or final impressions to maintain conciseness and ensure that the main message is easily understood by the reader (Lee et al., 2013). For example, a report may state "congestive heart failure" without mentioning sub-findings that are commonly associated with the condition, such as "consolidation" or "cardiomegaly". Consequently, it is often unclear whether unmentioned findings are truly absent or simply unrecorded, even though knowing this distinction is crucial, since misinterpreting unrecorded evidence as true absence can systematically bias the data and distort the inferred diagnostic reasoning (Gärtner et al., 2020; Lee et al., 2013; Gunderman, 2009; Berlin, 2000).

In other words, implicit uncertainty does not stem from linguistic ambiguity, but from the selective and incomplete nature of clinical reporting (Turner et al., 2021) – what evidence is stated, abstracted, or left unstated within the report. Disentangling these possibilities requires contextual understanding of diagnostic logic and domain expertise, making implicit uncertainty difficult to model and largely unaddressed in prior research. Understanding and modeling these uncertainties is not only critical for structuring radiology reports but also for developing AI systems that faithfully capture the reasoning process of radiologists. By explicitly representing both what is uncertain and what is implied, such structured resources enable more reliable training and evaluation of medical AI models in uncertainty-aware report generation and interpretation.

To address this gap, we aim to fill in the missing findings that are not explicitly mentioned but are implied by the stated diagnoses when structuring radiology reports, as is shown in Figure 1. We **construct expert-defined diagnostic pathways for 14 common CXR conditions**, capturing characteristic sub-findings typically observed with high likelihood (>80%). These pathways are then used to enrich the original reports by deterministically adding sub-findings that support explicitly mentioned diagnoses, with diagnostic certainty (derived from our explicit uncertainty extraction pipeline) propagated to each added finding. This approach results in expanded structured reports that more accurately reflect the radiologist's underlying reasoning.

In summary, our contributions are as follows:

- **Quantifying explicit uncertainty**: We introduce a comprehensive framework to estimate the probability of findings in radiology reports, taking into account hedging phrases and their sentence-level context. As part of this framework, we publish an expert-validated reference ranking of common hedging phrases.
- **Addressing implicit uncertainty**: We present the first framework to model implicit uncertainty in radiology reports by releasing diagnostic pathways for 14 common CXR diagnoses and integrating them into a rule-based framework. This framework reconstructs omitted diagnostic evidence by inferring the sub-findings that support each diagnosis.
- **Releasing LUNGUAGE^{++}**, which extends the LUNGUAGE dataset of structured radiology reports (Moon et al., 2026) by incorporating our techniques for capturing explicit and implicit uncertainty.

Code can be found at our Github repository[1], while LUNGUAGE^{++} and related resources will be made available via Physionet.

2. Lunguage Dataset

We demonstrate our methods on the LUNGUAGE dataset (Moon et al., 2026), a benchmark dataset containing 1,473 annotated CXR reports from the MIMIC-CXR dataset (Johnson et al., 2019b). Each report has been structured into fine-grained *(finding, relation, attribute)* triplets, where findings represent core clinical concepts (e.g. "opacity", "pneumonia"), and relations specify their contextual links

[1] github.com/prabaey/lunguage_uncertainty

(e.g., “location”, “severity”) with corresponding attributes (e.g., “right lower lobe”, “moderate”). We use the LUNGUAGE dataset because of its extensive granularity in the included findings, relations and attribute types. Furthermore, each annotated finding includes a binary label (*tentative* and *definitive*) quantifying the confidence expressed by the radiologist – hedging phrases (*e.g.,* “suggests”, “cannot exclude”) lead to a *tentative* label, while findings that lack such phrases are labeled *definitive*.

In this work, we represent each report $\mathcal{R}$ as a structured set of $n_{\mathcal{R}}$ findings extracted from $m_{\mathcal{R}}$ sentences:

$$\mathcal{R} = (\{(f_i, s_i, c_i, a_i)\}_{i=1}^{n_{\mathcal{R}}}, \{t_j\}_{j=1}^{m_{\mathcal{R}}}), \quad (1)$$

Here, t_j denotes the textual form of each **sentence**, f_i indicates the **finding**, s_i indicates the **status** (*positive* if the finding is present and *negative* if it is absent), c_i indicates the **certainty** (*definitive* or *tentative*), and a_i represents the **attributes** (e.g., location, morphology) associated with each finding. In LUNGUAGE, there are 14,049 such structured findings. In our study, we exclude sentences from the *history* section of the reports, as this section primarily contains information about patient symptoms or prior clinical records, whereas our analysis focuses on the *findings* and *impression* sections that convey diagnostic observations and reasoning.

3. Explicit uncertainty

We assume we have a report $\mathcal{R}$, where the sentence t_j expresses some uncertainty about finding f_i, in other words $c_i = tentative$. Across all reports in LUNGUAGE, this leaves us with a subset of 2,066 tentative finding-sentence pairs. Our goal is to quantify this uncertainty by assigning a probability p_i between 0 (finding certainly absent) and 1 (finding certainly present). In the remainder of this section, we ignore the explicit status s_i, as the target probability p_i inherently captures the presence or absence of the finding. Furthermore, we disregard all findings where $c_i = definitive$, as there is no uncertainty in this case. From this point onward, a *sentence* refers to the complete text t_j, which contains a target *finding* f_i (e.g. “pneumonia”) and one or more corresponding *(hedging) phrase(s)* (e.g., “possible”), expressing the degree of uncertainty associated with f_i.

Previous attempts to map hedging phrases to probabilities relied on expert ratings of individual phrases, an approach shown to be unreliable due to inconsistency across experts (Hobby et al., 2000; Shinagare et al., 2019). Furthermore, the context of the sentence beyond individual phrases should be taken into account: “probably pneumonia” conveys a different certainty than “probably pneumonia given patient history”, even though both use the phrase “probably”. To address these limitations, we adopt a different strategy, which relies on large language models (LLMs) to perform in-context pairwise comparisons between sentences conveying uncertainty. An overview of our framework is shown in Figure 2. We now describe each step of the process in detail. Additional details can be found in Appendix A.

3.1. Extracting a vocabulary of common hedging phrases

For each finding-sentence pair, we automatically extract the hedging phrases that convey uncertainty about the finding f_i. We do this by prompting Gemini (Comanici et al., 2025). The main part of the prompt is shown in Listing 1, while the full prompt contains a system message, ten in-context examples and additional instructions (see Appendix A.1). The prompt specifies the finding f_i to avoid extraction of hedging phrases from the sentence which have nothing to do with that particular finding. We include in our vocabulary all hedging phrases that were extracted ten times or more, resulting in a vocabulary of 42 hedging phrases. Each phrase is associated with a list of finding-sentence pairs where it was extracted. The five most common extracted hedging phrases include *or* (373 times), *likely* (239 times), *may* (215 times), *suggesting* (74 times), and *cannot be excluded* (71 times); the full vocabulary is found in Appendix A.1.

Listing 1: Prompt for hedging phrase extraction

```
Your task is to identify and extract only the words
    or phrases in the sentence that express
    uncertainty specifically about the given
    finding.
```

3.2. Building a reference ranking of hedging phrases

We construct a reference ranking of the 42 common hedging phrases, where the top rank corresponds to a probability near 1 and the bottom rank near 0.

TrueSkill We draw inspiration from ranking systems in competitive gaming, where player skill is inferred from the outcomes of matches. Specifically, we employ the TrueSkill algorithm (Herbrich et al., 2006), a Bayesian rating system that updates each item’s mean skill level μ based on pairwise comparisons with other items. This way, TrueSkill efficiently infers a global ranking that reflects the relative skill level μ of each item. In our case, the items are hedging phrases.

LLM as a judge To obtain a reliable and well-calibrated ranking, a large number of pairwise comparisons between phrases is required. We therefore leverage LLMs to perform these comparisons automatically, employing three general-purpose LLMs (Gemini (Comanici et al., 2025),

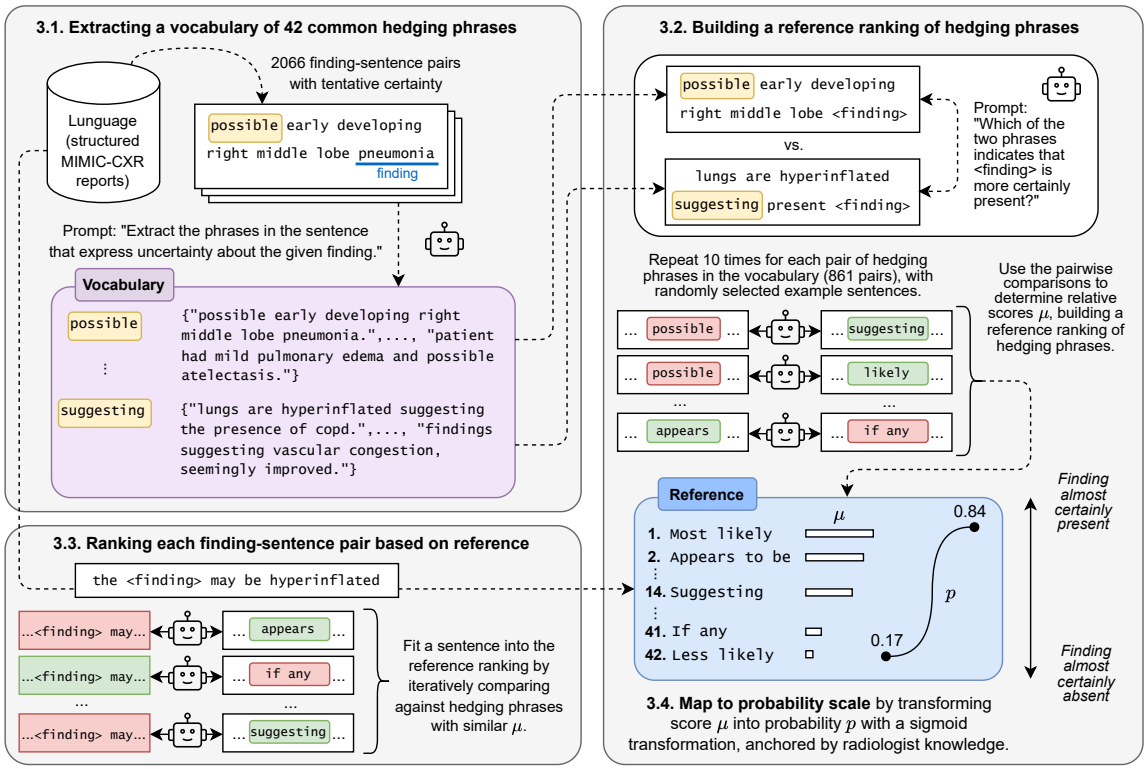


Figure 2: Strategy for assigning probabilities to finding-sentence pairs with tentative certainty in the LUNGUAGE dataset: We first build a vocabulary of common hedging phrases and the sentences in which these are used (Section 3.1). Next, we leverage LLMs to construct a reference ranking of these phrases, by performing pairwise comparisons of examples sentences (Section 3.2). Each finding-sentence pair is then compared against this reference (Section 3.3) and is finally mapped to a probability (Section 3.4). This approach ensures that the probability assigned to each finding reflects not only the hedging phrase itself but also the broader context in which it appears.

GPT-4o (Hurst et al., 2024), and Claude (Anthropic, 2025)) and one domain-specific medical LLM (MedGemma (Sellergren et al., 2025)).[2] Since the preferred phrase can vary depending on the sentence context, we conduct ten comparisons for each pair of the 42 hedging phrases, where in each comparison, we randomly sample a sentence in which the phrase occurs. We ask each LLM to identify which phrase conveys greater certainty about the presence of a finding, ensuring that such phrases will eventually receive a higher μ. The main portion of the prompt is shown in Listing 2; the complete prompt additionally includes a system instruction, detailed task description, and three in-context examples (Appendix A.2). To ensure a more neutral comparison, the referenced finding f_i within each sentence is masked as <*finding*>.

Listing 2: Prompt for hedging phrase comparison

```
You will be given two sentences from radiology
    reports. Each sentence contains a placeholder <
    finding>, which represents a medical
    observation. Your task is to identify which
    sentence expresses a higher degree of certainty
     that the finding is present.
```

[2]To deal with the sensitive nature of the reports in LUNGUAGE we (i) ran a HIPAA-compliant GPT-4o model provided by Azure, (ii) revoked data retention rights for Gemini and Claude, and (iii) ran MedGemma locally.

Expert	Ref.	Gemini	GPT-4o	Claude	MedGem.
Radiologist	0.80	0.80	0.82	0.72	0.84
Radiologist	0.80	0.76	0.82	0.76	0.80
Internist	0.86	0.86	0.88	0.82	0.90
Oncologist	0.72	0.68	0.74	0.76	0.72
GP	0.66	0.70	0.68	0.66	0.62
GP	0.82	0.82	0.84	0.78	0.82
Average	0.78	0.77	0.80	0.75	0.77

Table 1: Expert agreement with the reference ranking and individual LLMs. We define agreement as the proportion of the 50 phrase pairs where expert and model judgments are concordant. GP = General Practitioner.

This procedure yields 4×8610 pairwise comparisons, each treated as an independent comparison by the TrueSkill algorithm. We execute TrueSkill 10 times with different random seeds that affect the order of matches, and then average the resulting μ values across runs. The final averaged scores produce a stable reference ranking of hedging phrases, which is shown in Figure 3. Appendix A.2 assesses the robustness of this reference ranking and explores inter-LLM agreement across the set of comparisons.

Expert evaluation To validate the final reference ranking, we conducted an expert evaluation study. We recruited two expert *writers* (radiologists) and four expert *readers* (internist, oncologist, and two general practitioners). Each participant was pre-

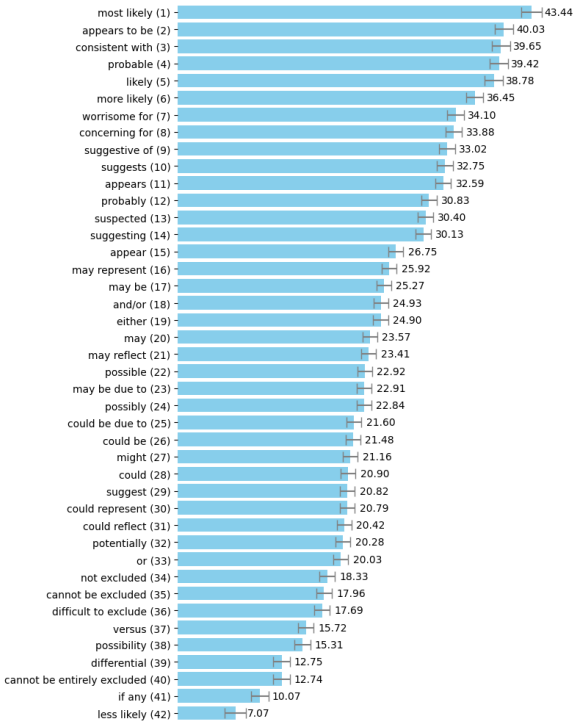


Figure 3: Reference ranking of the 42 common hedging phrases in our vocabulary. The mean skill level μ for each phrase is shown on the right, with the confidence σ represented by the error bars. Phrases at the top of the ranking correspond to a high likelihood that the finding is present, while phrases at the bottom correspond to a high likelihood that the finding is absent.

sented with 50 pairs of hedging phrases, with five example sentences per phrase, all randomly sampled from our vocabulary. Participants were asked to select the phrase that conveyed a higher degree of certainty that the finding was present. All participants evaluated the same set of 50 phrase pairs. See Appendix A.2 for additional details.

Table 1 presents the agreement between each expert and the reference ranking. Agreement is defined as the proportion of phrase pairs (out of 50) for which the expert's relative ordering of the two phrases matches that of the reference. In addition, we assess the agreement between each expert and each LLM. Since each LLM evaluated every phrase pair ten times using different example sentences, we first derive a consensus decision for each pair through majority voting, resolving ties at random.

We see that the experts agree well with both the reference ranking and the individual LLMs. Furthermore, the inter-expert Fleiss' κ is 0.72, indicating substantial agreement between experts. Among individual LLMs, GPT-4o shows the highest agreement with the experts, while Claude shows the lowest. Despite Claude's lower agreement on this limited subset of 50 pairs, we retain all LLMs in the reference ranking, to leverage the diversity of judgments across models, thereby improving the robustness of the TrueSkill-based ranking. Moreover, the expert evaluation covers only a small fraction of the full dataset and uses a different evaluation protocol than the sentence-level LLM comparisons, so occasional disagreements by a single LLM are not sufficient reason for exclusion.

3.3. Ranking each finding-sentence pair based on the reference ranking

With the reference ranking established, we fit each of the 2,066 tentative finding-sentence pairs from the LUNGUAGE dataset into it. Note that we cannot simply use the hedging phrases extracted from the sentence to assign a rank directly, because (i) some phrases are too rare to appear in the reference vocabulary, and (ii) the context in which a phrase occurs can significantly alter its implied certainty. Here, we once again draw inspiration from competitive gaming: when a new player enters a game, TrueSkill quickly estimates their rank by identifying existing players with similar skill levels μ. This is done by selecting opponents with the highest *draw probability*, iteratively playing those games, and updating the skill level of the new player based on the outcome of each game (Herbrich et al., 2006).

To fit a finding–sentence pair into the reference ranking, we initialize the target sentence's parameter μ to the default value of 25 and compute its draw probability against all opponent phrases in the reference ranking. The phrase with the highest draw probability is selected, and a corresponding sentence is randomly sampled from the vocabulary. Using the prompt from Listing 2, the target and opponent sentence are compared. During the first $K = 10$ iterations, all four LLMs perform the comparison; thereafter, one LLM is selected at random to reduce cost. Based on the outcome, the TrueSkill algorithm updates μ for the target sentence. Draw probabilities are then recomputed, and the next opponent is selected accordingly, with each opponent limited to $N = 5$ comparisons. The procedure terminates once the target sentence's rank remains stable for ten consecutive steps, or after 100 iterations, whichever occurs first. Through hyperparameter tuning, we set K to 10 and N to 5. Applying our algorithm to the 2,066 tentative finding-sentence pairs in LUNGUAGE incurred a total cost of $92.16, averaging $0.045/pair. The full algorithm, including experiments to validate our opponent selection strategy using *draw probability*, can be found in Appendix A.3.

3.4. Map to probability scale

In the final step, we map the TrueSkill score μ, which determines the position of each finding-sentence pair in the ranking, to a probability $p \in [0, 1]$. This transformation is achieved using the sigmoid function $p = 1/(1 + e^{-\alpha(\mu-\mu_0)})$, where α

controls the steepness of the curve and μ_0 is the inflection point corresponding to a probability of 0.5.

We determine α and μ_0 using two anchor points: the desired probability p_{bottom} for the phrase *less likely* (lowest-ranked, $\mu = 7.07$), and p_{top} for the phrase *most likely* (highest-ranked, $\mu = 43.44$). To obtain these anchors, two radiologists independently reviewed ten example sentences per phrase and assigned a probability between 0 ("certainly absent") and 1 ("certainly present"); see Appendix A.4 for full instructions. Averaged across radiologists and examples, the resulting values were $p_{bottom} = 0.170$ (95% CI: $[0.156, 0.185]$) and $p_{top} = 0.839$ (95% CI: $[0.818, 0.860]$). These anchors yield $\alpha = 0.089$ and $\mu_0 = 24.89$.

After applying this mapping, the average probability for the 2,066 tentative finding-sentence pairs in LUNGUAGE is 0.459, with a standard deviation of 0.185, a maximum of 0.892, and a minimum of 0.102. Table 2 summarizes statistics for common CXR findings in the LUNGUAGE dataset. For each finding, we report proportions of positive and negative cases with definitive certainty, and the mean, standard deviation, minimum, and maximum probability for cases with tentative certainty.

4. Implicit uncertainty

Radiology reports often encode diagnostic reasoning implicitly (Gunderman, 2009; Lee et al., 2013; Berlin, 2000; Gärtner et al., 2020), documenting high-level diagnoses while omitting intermediate supporting findings. Although such abstraction enhances efficiency, it introduces ambiguity arising from unmentioned findings. To resolve this, we propose a **Pathway Expansion Framework** that reconstructs omitted diagnostic evidence by systematically expanding structured findings along predefined **diagnostic pathways**. This framework yields LUNGUAGE^{++}, an expanded version of LUNGUAGE, integrating explicit and implicit reasoning into a unified structured representation (Figure 4).

4.1. Diagnostic Pathway Construction

We formalize radiologists' implicit diagnostic reasoning into explicit, machine-friendly formats that describe how each diagnosis decomposes into its characteristic radiographic findings. To preserve interpretability and clinical fidelity, we established three principles through expert consensus. (i) **Exclusive mutual independence** ensures that each pathway differs from others by at least one defining piece of evidence. (ii) **Specificity** ensures that observations directly contributing to diagnostic differentiation are included, emphasizing findings that are concrete and fine-grained. (iii) **High-certainty features** retain only clinically consistent and reliable findings to minimize ambiguity and preserve clarity. Pathways were then constructed through a two-stage expert-in-the-loop process grounded in these principles. A radiologist and an AI researcher first defined subfindings for 14 common CXR diagnoses, referencing established radiological interpretation principles (Goodman, 2014; Webb and Higgins, 2011) to identify diagnostic criteria, imaging patterns, and hierarchical relations among findings. These preliminary pathways were subsequently reviewed and refined through consensus by another radiologist and an oncologist.

Pathways are represented as Directed Acyclic Graphs (DAGs) (Pearl, 2009), capturing how higher-level diagnoses require lower-level manifestations within hierarchical reasoning. An example of a DAG corresponding to (part of) the diagnostic pathways for *congestive heart failure* (*CHF*), *pulmonary edema*, *pleural effusion*, and *cardiomegaly* is shown in Figure 4. Each node corresponds to an observable finding, and directed edges (u, v) represent diagnostic relations $p_{path}(v|u)$ established through radiologist consensus, indicating that the sub-finding v is typically observed when its parent u is present, with likelihood greater than 0.8. This top-down structure models the cascading sequence of radiographic evidence that underlies diagnostic inference. Distinct pathways are defined for each diagnosis and further refined by imaging view and patient position, since these factors influence the findings one expects to observe. Table 3 reports statistical summaries of the diagnostic pathways, while the full set of pathways is provided in Appendix B.1 and on our Github repository.

4.2. Pathway Expansion Framework

Building upon the diagnostic pathways, we design a *Pathway Expansion Framework* that implements the top-down reasoning process of radiological interpretation to fill in missing sub-findings. Given structured findings from LUNGUAGE (Moon et al., 2026), the framework reconstructs omitted intermediate findings by expanding each diagnosis along its predefined diagnostic pathway, following the recursive procedure detailed in Algorithm 1. It consists of four stages, each of which ensure that reconstructed structures remain clinically faithful and logically consistent.

1. Finding Deduplication To ensure that subsequent reasoning operates on a consistent and non-redundant set of findings, overlapping or synonymous mentions referring to the same diagnostic concept at the same location (e.g., "opacity in the basal segment of the lung" and "lung base opacity") are identified and merged within each report. Pairwise cosine similarity is computed between clinical embeddings (BioLORD (Remy et al., 2023)) of all report findings, where each finding is linearized into a phrase combining its *"entity, location, attributes"* following the LUNGUAGESCORE (Moon et al., 2026)

	Certainty level		Definitive status		Tentative probability			
Finding	**Definitive**	**Tentative**	**Positive**	**Negative**	**Avg.**	**Std.**	**Min.**	**Max.**
Pleural effusion	1165 (86.2%)	186 (13.8%)	497 (42.7%)	668 (57.3%)	0.465	0.201	0.102	0.862
Pneumothorax	897 (98.8%)	11 (1.2%)	43 (4.8%)	854 (95.2%)	0.480	0.151	0.269	0.709
Atelectasis	419 (62.2%)	255 (37.8%)	408 (97.4%)	11 (2.6%)	0.536	0.156	0.145	0.876
Pulmonary edema	516 (83.5%)	102 (16.5%)	264 (51.3%)	252 (48.7%)	0.479	0.186	0.102	0.841
Consolidation	406 (84.1%)	77 (15.9%)	90 (22.2%)	316 (77.8%)	0.447	0.182	0.113	0.808
Pneumonia	225 (51.3%)	214 (48.7%)	32 (14.2%)	193 (85.8%)	0.423	0.221	0.102	0.872
Cardiomegaly	313 (97.8%)	7 (2.2%)	308 (98.4%)	5 (1.6%)	0.518	0.141	0.365	0.706
Congestive heart failure	49 (89.1%)	6 (10.9%)	23 (46.9%)	26 (53.1%)	0.591	0.248	0.102	0.765
Emphysema	40 (80.0%)	10 (20.0%)	40 (100.0%)	0 (0.0%)	0.518	0.134	0.314	0.660
COPD	18 (60.0%)	12 (40.0%)	18 (100.0%)	0 (0.0%)	0.537	0.050	0.453	0.594
Fracture	12 (85.7%)	2 (14.3%)	12 (100.0%)	0 (0.0%)	0.219	0.165	0.102	0.336
Lung cancer	6 (46.2%)	7 (53.8%)	6 (100.0%)	0 (0.0%)	0.531	0.171	0.249	0.662
Tuberculosis	3 (60.0%)	2 (40.0%)	2 (66.7%)	1 (33.3%)	0.413	0.051	0.377	0.449
Bronchitis	1 (50.0%)	1 (50%)	1 (100.0%)	0 (0.0%)	0.475	0.000	0.475	0.475

Table 2: Certainty statistics for common CXR findings in the LUNGUAGE dataset. For each finding, we collect the instances in LUNGUAGE where f_i is mapped to the finding through its various lexical and contextual variants, following the approach described in Section 4.2 (Finding Deduplication).

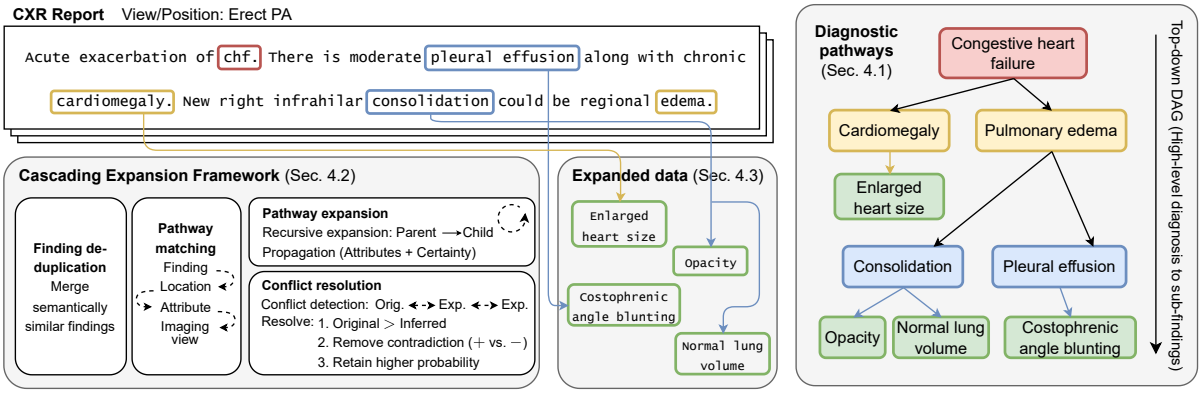


Figure 4: Overview of the Pathway Expansion Framework. The framework expands structured findings from LUNGUAGE along diagnostic pathways (Section 4.1) to reconstruct omitted diagnostic evidence. It comprises four stages—finding deduplication, pathway matching, pathway expansion, and conflict resolution (Section 4.2)—that jointly ensure semantic coherence and clinical validity. The resulting representation connects high-level diagnoses with their underlying evidence, forming LUNGUAGE^{++}, which is further analyzed in Section 4.3.

formulation. Pairs exceeding a similarity threshold of 0.9 are merged into a single finding, while manually defined *blacklist pairs* (e.g., left vs. right) are excluded to prevent erroneous merges of semantically distinct findings. This precision-oriented threshold and manual blacklist correction minimize erroneous merges while ensuring that each finding is represented only once before expansion. The detailed blacklist pairs are provided in Appendix B.2.1.

2. Pathway Matching After deduplication, the framework aligns each finding in LUNGUAGE with its most appropriate diagnostic pathway (Section 4.1) through a sequential process based on its *finding*, *location*, *attributes*, and *imaging view*. All terms are first normalized using the LUNGUAGE vocabulary to ensure consistent terminology. The process begins with *finding-level matching*, which identifies the core diagnostic concept but may yield ambiguous interpretations. For instance, "fracture" can denote either a medical device fracture (e.g., pacemaker lead) or a thoracic skeletal fracture (e.g., rib, spine). Such ambiguity is resolved through *location matching*, which determines whether the finding pertains to an anatomical structure within the thorax or not, thereby selecting the appropriate pathway (e.g., device vs. skeletal fracture). Subsequently, *attribute matching* refines the alignment by specifying diagnostic stages such as "acute", "chronic", or "healed". Finally, *imaging view* and *patient orientation* (e.g., "PA, erect") define the view-specific pathway variant, ensuring that inferred findings remain anatomically observable within the imaging context. For the detailed process, see Appendix B.2.2.

3. Pathway Expansion Building on the matched diagnostic pathways, the framework recursively expands structured findings in LUNGUAGE by inferring omitted sub-findings through parent–child relations defined in each pathway (Algorithm 1). The expan-

	Diagnostic Pathway Structure				Expansion within LUNGUAGE†	
Diagnosis	**Variants**	**Pathways**	**Depth**	**Width**	**Expandable Finding**	**Inferred Sub-findings**
Pleural effusion	2	10	1	3	683 (4.9%)	+915 (6.6%)
Atelectasis	7	1	1	2	663 (4.7%)	+1,215 (8.6%)
Pulmonary edema	12	2	2	4	366 (2.6%)	+622 (4.4%)
Consolidation	4	1	1	2	167 (1.2%)	+801 (5.7%)
Pneumonia	22	3	2	3	249 (1.8%)	+311 (2.2%)
Cardiomegaly	1	1	1	1	315 (2.2%)	+340 (2.4%)
Congestive heart failure	5	1	3	6	29 (0.2%)	+79 (0.6%)
Pneumothorax	5	2	1	4	54 (0.4%)	+160 (1.1%)
Emphysema	6	2	1	3	50 (0.4%)	+98 (0.7%)
COPD	3	1	2	5	30 (0.2%)	+88 (0.6%)
Fracture	15	4	1	5	14 (0.1%)	+25 (0.2%)
Lung cancer	7	1	1	1	13 (0.1%)	+12 (0.1%)
Tuberculosis	5	2	1	3	4 (0.0%)	+2 (0.0%)
Bronchitis	4	2	1	3	2 (0.0%)	+2 (0.0%)
Total / Avg.	**98**	**33**	**1.4**	**3.2**	**2,639 (18.7%)**	**+4,761 (33.9%)**

Table 3: Statistics of Diagnostic Pathways and LUNGUAGE++. The left panel summarizes the hierarchical structure of 14 expert-defined diagnostic pathways (98 variants, 33 unique DAGs). The right panel quantifies their application within the 14,049 structured findings in LUNGUAGE. A total of **2,636 findings (18.7%)** were eligible for expansion (i.e., directly aligned with a pathway node) serving as anchors for hierarchical reasoning. From these anchors, the framework inferred **4,761 additional sub-findings (+33.9%)**, representing previously implicit but clinically consistent observations reconstructed from the pathway hierarchy. †*Percentages are relative to the 14,049 original findings in LUNGUAGE. "+" denotes additional findings inferred during expansion.*

Algorithm 1 Recursive Finding Expansion

Input: Parent node f_i in pathway $\mathcal{P}$ with s_i, c_i, p_i, a_i
Output: Set $\mathcal{E}$ of all sub-findings expanded to leaf nodes

1: **function** EXPAND($f_i, s_i, c_i, p_i, a_i, \mathcal{P}$)
2: $\mathcal{E} \leftarrow \emptyset$
3: **if** $s_i \neq$ *positive* **then**
4: **return** $\emptyset$ *// Expand only positive findings*
5: **end if**
6: **for** f_j where $(f_i, f_j) \in \mathcal{P}$ **do** *// For each child f_j in $\mathcal{P}$*
7: $(s_j, c_j, p_j) \leftarrow (s_i, c_i, p_i)$ *// Inherit f_i properties to f_j*
8: **if** pathway $(f_i \rightarrow f_j)$ defines refinements **then**
9: $a_j \leftarrow a_i \cup$ attributes specified for f_j
10: **else**
11: $a_j \leftarrow a_i$ *// Inherit as-is*
12: **end if**
13: *// Add expanded child f_j to $\mathcal{E}$*
14: $\mathcal{E} \leftarrow \mathcal{E} \cup \{(f_j, s_j, c_j, p_j, a_j)\}$
15: **if** $\mathrm{HasChildren}(f_j, \mathcal{P})$ **then** *// recurse for f_j*
16: $\mathcal{E} \leftarrow \mathcal{E} \cup$ EXPAND($f_j, s_j, c_j, p_j, a_j, \mathcal{P}$)
17: **end if**
18: **end for**
19: **return** $\mathcal{E}$
20: **end function**

sion begins with a parent finding f_i, which corresponds to one of the 14 diagnoses in the diagnostic pathways $\mathcal{P}$. Each finding is represented by its diagnostic status s_i (*positive* or *negative*), certainty c_i (*definite* or *tentative*), probability p_i (quantified in Section 3), and attributes a_i (e.g., location, morphology). During expansion, every child node in $\mathcal{P}$ inherits all diagnostic properties from its parent—s_i, c_i, p_i, and a_i—ensuring that each descendant finding reflects its parent's context. For example, if f_i is labeled as *positive, tentative* with $p_i = 0.6$, all inferred sub-findings inherit the same label and probability. However, findings with $s_i =$ *negative* are not expanded (line 3 in Algorithm 1), since the pathway $\mathcal{P}$ is not reversible; for instance, "no pneumonia" does not imply "no opacity." This inheritance (lines 7–11) reflects the propagation of diagnostic properties and the pathway-specific modification of attributes. Comprehensive expansion is achieved through recursive chaining (line 16): if a child finding f_j corresponds to one of the 14 diagnoses in $\mathcal{P}$, it becomes a new parent node and triggers further expansion until reaching leaf nodes. For example, when "pulmonary edema" acts as a parent node, it expands to "consolidation," which in turn recursively expands to its subfinding "opacity." This recursive expansion ensures that all findings along the diagnostic pathway maintain consistent uncertainty estimates and coherent hierarchical relations.

4. Conflict Resolution Although pathway relations capture generally valid diagnostic dependencies (Section 4.1), clinical exceptions and inconsistent uncertain expression can lead to conflicting findings during expansion. These conflicts typically arise when multiple pathways generate overlapping findings, leading to discrepancies in diagnostic status (positive vs. negative) or certainty (definitive vs. tentative). Conflicts were classified as: (i) *Original vs. Expansion* — discrepancies between the original report and expansion (e.g., a resolved "pulmonary edema" reappearing from the CHF pathway); and (ii) *Expansion vs. Expansion* — contradictions among expansion (e.g., "volume loss" from atelectasis vs. "normal volume" from consolidation).

To ensure clinical plausibility, all potential conflicts were reviewed with radiologist input and resolved through a rule-based consistency protocol. Resolution followed a sequential protocol: (1) original findings take precedence to preserve factual accuracy; (2) contradictory entries (e.g., positive vs. negative) are removed; and (3) among uncertain cases, the instance with higher probability p_i is retained. This process ensures logical coherence and clinical consistency across all expanded findings. The detailed procedure is presented in Appendix B.2.3.

4.3. Analysis of LUNGUAGE^{++}

We analyze the diagnostic pathways constructed in Section 4.1 and the expanded dataset generated by the framework in Section 4.2. This analysis quantifies the extent to which the framework reconstructs implicit diagnostic reasoning, propagates uncertainty, and maintains report-level coherence. Table 3 reports statistics for the full set of diagnostic pathways.

Diagnostic Pathway Diversity Pathway diversity arises from variations in *imaging view*, *finding*, *location*, and *attributes*, which inherently capture differences in disease stage (e.g., acute vs. chronic). Across **14 diagnostic categories** (e.g., pneumonia), we identified **98 disease variants** (e.g., hospital-acquired pneumonia) organized into **33 unique diagnostic pathways** (e.g., pneumonia vs. lobar pneumonia), as shown in Table 3. Each pathway is characterized by its **depth**, representing the maximum hierarchical expansion depth, and its **width**, denoting the number of leaf-level findings at the terminal layer. On average, pathways exhibit a depth of 1.4 and a width of 3.2, indicating that most reasoning chains consist of one to three inferential layers. Among them, *pleural effusion* shows the greatest diversity (**10 pathways**) due to view-dependent fluid distribution, whereas *congestive heart failure* exhibits the deepest hierarchy (depth 3), cascading through cardiomegaly, pulmonary edema, and pleural effusion (Figure 4). In contrast, simpler findings (e.g., *cardiomegaly*) follow single-layer mappings (depth 1, width 1), directly linking diagnosis to observation.

Expansion Coverage As shown in Table 3, applying the Pathway Expansion Framework to 14,049 structured findings in LUNGUAGE identified **2,639 findings (18.7%)** that aligned with predefined diagnostic pathways, from which **4,761 additional sub-findings (+33.9%)** were inferred. This shows that roughly one in five findings contained implicit diagnostic structure recoverable through pathways. Expansion was dominated by interdependent categories such as *atelectasis* (**+1,215**), *pleural effusion* (**+915**), and *consolidation* (**+801**), which together accounted for more than half of all inferred findings. These conditions frequently co-occur or appear in hierarchical cascades naturally producing richer expansions, e.g. effusion with edema or atelectasis.

Conflict Resolution To assess the stability of the expanded structures, we further examined conflicts that arose during expansion. Across the entire LUNGUAGE dataset, such conflicts were rare, appearing in approximately **3.2%** of cases overall: **0.9%** arose between original and expanded findings (64.8% status, 35.2% certainty), and **2.3%** between expansions (22.8% status, 77.2% certainty). After conflict resolution, all remaining inconsistencies were resolved, confirming that the expanded dataset maintains both logical consistency and clinical validity. Detailed analyses are presented in Appendix B.2.3.

5. Conclusion

In this work, we present the first systematic approach to addressing both explicit and implicit uncertainty in radiology reports. We rigorously quantify the degree of **explicit uncertainty** of findings in CXR reports, through an LLM-based automated framework. While demonstrated on the LUNGUAGE dataset, this framework can be applied to any CXR report corpus, enabling the enrichment of widely used benchmarks such as CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019a) (of which LUNGUAGE^{++} is only a subset) with continuous uncertainty measures.

In parallel, we expose and address **implicit uncertainty** in radiology reports arising from omitted elements of diagnostic reasoning. We introduce a rule-based expansion framework based on expert-defined diagnostic pathways for 14 common CXR diagnoses, which add characteristic sub-findings that may have been left unstated in the original reports. The diagnostic pathways can easily be reused for other tasks, while the expansion framework is applicable to any dataset that has been structured using the LUNGUAGE framework proposed by Moon et al. (2026).

Together with our reusable frameworks, we release LUNGUAGE^{++}, a benchmark dataset of structured radiology reports that includes continuous probabilities for all extracted findings and pathway-based expansions that expose previously omitted findings. This enriched resource supports a range of future research directions, including training uncertainty-aware CXR image classifiers, guiding vision-language models toward uncertainty-aware report interpretation and generation, and studying how diagnostic uncertainty influences downstream clinical outcomes.

Limitations

For the **explicit uncertainty framework**, a key limitation lies in our reliance on LLM-based pairwise comparisons to construct the reference ranking of hedging phrases. While this approach offers the advantage of scalability (enabling 8,610 comparisons across four LLMs, far beyond what would be feasible with human raters) it also introduces dependency on model behavior. Moreover, our mapping from skill level μ to probability p is based on assessments from only two radiologists; incorporating a larger expert pool would better capture the variability in how uncertainty is interpreted, as noted in prior studies (Hobby et al., 2000; Shinagare et al., 2019). Finally, the current framework incurs nontrivial costs when assigning probabilities to new sentences, due to repeated LLM comparisons with the reference ranking. Future work could mitigate this by training a lightweight, locally hosted model, fine-tuned on our high-quality LLM comparison data.

For the **implicit uncertainty framework**, the limitations include both structural and clinical aspects. First, the diagnostic pathways are constructed as a top-down Directed Acyclic Graph that maps high-level diagnoses to lower-level sub-findings. While this design captures hierarchical reasoning, it remains unidirectional and cannot represent bottom-up or cyclic dependencies that naturally arise in clinical reasoning – such as when multiple findings interact or reinforce one another to revise a diagnosis. Future work could extend this structure into a bidirectional or dynamically learnable graph representation (e.g., a Bayesian network) that allows reasoning to flow in both directions, thereby capturing the iterative nature of diagnostic interpretation.

Second, when expanding findings along diagnostic pathways, we currently assign the same probability to all child nodes. For instance, if *congestive heart failure (CHF)* has a probability of 0.8, its inferred sub-findings—*cardiomegaly* and *pulmonary edema*—are each assigned the same value ($p = 0.8$), even though in practice, one may be more certain (e.g., $p = 0.9$) while the other less so (e.g., $p = 0.6$). This simplification overlooks interdependence and uncertainty calibration among findings, which future work could address by modeling probabilistic propagation that accounts for relative diagnostic confidence.

Lastly, despite being defined through high-probability relations ($p(v|u) > 0.8$), the pathways inherently encompass clinical exceptions. Even strongly associated findings may not hold under atypical imaging conditions or in the presence of comorbidities, occasionally causing conflicts between pathway-inferred and explicitly reported findings. These exceptions reflect the inherent variability of radiological practice rather than modeling errors, but they underscore the need for incorporating image-level verification and cross-modal reasoning to refine the framework. Future iterations could leverage visual grounding to reconcile such exceptions, ensuring that inferred reasoning aligns more closely with true clinical evidence.

Data and Code Availability

LUNGUAGE^{++} and related resources will be made available via Physionet. Code can be found at our Github repository: github.com/prabaey/lunguage_uncertainty.

Ethics Statement

This research made use of and expanded the LUNGUAGE dataset (Moon et al., 2026). This dataset is derived from MIMIC-CXR (Johnson et al., 2019b), a public dataset of chest radiographs and free-text radiology reports. As this dataset contains sensitive patient information, we will follow Physionet's guidelines by publishing LUNGUAGE^{++} under the same agreement as the source data. Furthermore, to avoid passing sensitive patient data to public LLM APIs, we followed Physionet's recommendations by (i) running a HIPAA-compliant GPT-4o model provided by Azure, (ii) revoking data retention rights for Gemini and Claude, and (iii) running MedGemma locally.

Acknowledgments

Paloma Rabaey's research is funded by the Research Foundation Flanders (FWO Vlaanderen) with grant number 1170124N. Additional FWO funding was provided in the form of a travel grant, with grant number V414025N. This research also received funding from the Flemish government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.RS-2019-II190075, No.RS-2025-02304967), the Korea Health Industry Development Institute (KHIDI) grant (No.RS-2025-02213750), and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945), funded by the Korean government (MSIT, MOHW).

The authors would like to thank Géraldine Deberdt, Stefan Heytens and Johan Decruyenaere for participating in the expert evaluation study. Additionally, we are grateful to Stefan Heytens and Johan Decruyenaere for their input on the conceptual design of the expert evaluation study and sharing their view on uncertainty in medical reporting.

Bibliographical References

Anthropic. 2025. Claude sonnet. https://www.anthropic.com/transparency/model-report.

Leonard Berlin. 2000. Pitfalls of the vague radiology report. *American Journal of Roentgenology*, 174(6):1511–1518.

Michael A Bruno, Jonelle Petscavage-Thomas, and Hani H Abujudeh. 2017. Communicating uncertainty in the radiology report. *American Journal of Roentgenology*, 209(5):1006–1008.

Andrew L Callen, Sara M Dupont, Adi Price, Ben Laguna, David McCoy, Bao Do, Jason Talbott, Marc Kohli, and Jared Narvid. 2020. Between always and never: evaluating uncertainty in radiology reports using natural language processing. *Journal of Digital Imaging*, 33(5):1194–1201.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Julia Gärtner, Pascal O Berberat, Martina Kadmon, and Sigrid Harendza. 2020. Implicit expression of uncertainty–suggestion of an empirically derived framework. *BMC Medical Education*, 20(1):83.

Lawrence R Goodman. 2014. *Felson's principles of chest roentgenology, a programmed text*. Elsevier Health Sciences.

Richard B Gunderman. 2009. Biases in radiologic reasoning. *American Journal of Roentgenology*, 192(3):561–564.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: a bayesian skill rating system. *Advances in neural information processing systems*, 19.

Jonathan L Hobby, BD Tom, C Todd, PW Bearcroft, and Adrian K Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73(873):999–1001.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019a. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101:215–220.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019b. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Sameer Khanna, Adam Dejl, Kibo Yoon, Steven QH Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. 2023. Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In *Machine learning for healthcare conference*, pages 381–402. PMLR.

Cindy S Lee, Paul G Nagy, Sallie J Weaver, and David E Newman-Toker. 2013. Cognitive and system factors contributing to diagnostic errors in radiology. *American Journal of Roentgenology*, 201(3):611–617.

Jong Hak Moon, Geon Choi, Paloma Rabaey, Min Gwam Kim, Hyuk Gi Hong, Jung Oh Lee, Hangyul Yoon, Eunwoo Doe, Jiyoun Kim, Harshita Sharma, Daniel Coelho de Castro, Javier Alvarez Valle, and Edward Choi. 2026. Lunguage: A Benchmark for Structured and Sequential Chest X-ray Interpretation. *PhysioNet*. Version 1.0.0.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Bruce I Reiner. 2018. Quantitative analysis of uncertainty in medical reporting: creating a standardized and objective methodology. *Journal of digital imaging*, 31(2):145–149.

François Remy, Kris Demuynck, and Thomas Demeester. 2023. Biolord-2023: Semantic textual representations fusing llm and clinical knowledge graph insights. *arXiv preprint arXiv:2311.16075*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Atul B Shinagare, Ronilda Lacson, Giles W Boland, Aijia Wang, Stuart G Silverman, William W Mayo-Smith, and Ramin Khorasani. 2019. Radiologist preferences, agreement, and variability in phrases used to convey diagnostic certainty in radiology reports. *Journal of the American College of Radiology*, 16(4):458–464.

Mark Turner, Julia Ive, and Sumithra Velupillai. 2021. Linguistic uncertainty in clinical nlp: A taxonomy, dataset and approach. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 129–141. Springer.

W Richard Webb and Charles B Higgins. 2011. *Thoracic imaging: pulmonary and cardiovascular radiology*. Lippincott Williams & Wilkins.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. 2021. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Mengliang Zhang, Xinyue Hu, Lin Gu, Tatsuya Harada, Kazuma Kobayashi, Ronald Summers, and Yingying Zhu. 2023. Cad-chest: Comprehensive annotation of diseases based on mimic-cxr radiology report. *PhysioNet*.

Appendix

A. Explicit Uncertainty

A.1. Vocabulary of hedging phrases

The full prompt that was used to extract hedging phrases for each finding-sentence pair in Language, is shown in Listing A1. We use this to prompt the Gemini model *gemini-2.5-flash* with the maximum number of output tokens set to 1100, where 1000 tokens were assigned as a thinking budget. *Temperature* and *top_p* parameters were set to their default values of 1. The total price to extract all hedging phrases was around $1.

Listing A1: Full prompt for hedging phrase extraction

```
SYSTEM: You are a radiologist who is given pairs of
     entities and sentences. Each entity appears in
     the corresponding sentence. Your task is to
     identify and extract only the words or phrases
     in the sentence that express uncertainty
     specifically about the given entity.

TASK: You are given pairs of entities and sentences.
      Each entity appears in the corresponding
     sentence. Your task is to identify and extract
     only the words or phrases in the sentence that
     express uncertainty specifically about the
     given entity.

Important notes:
- The sentence may mention multiple entities, but
     you should extract uncertainty clues only for
     the specified entity. Please refer to the
     examples below that show how you should handle
     such cases.
- Look for words or phrases that suggest uncertainty
     , speculation, approximation, possibility, or
     lack of definitiveness (e.g., "might", "
     possibly", "suggests", "appears to", "in some
     cases").
- Return a list of such uncertainty clues found in
     the sentence and relevant to the query entity.

Return your output as a list: ["<word or phrase 1>",
      "<word or phrase 2>", ...]
If there are no uncertainty clues related to the
     given entity, return an empty list.

Below are 10 examples, after which you must complete
      the task for an unseen query.

INPUT:
{
  "entity": "pulmonary edema",
  "sentence": "overall, however, there is a more
       focal airspace opacity in the left mid and
       lower lung, which may reflect asymmetric
       pulmonary edema or an infectious process,
       less likely atelectasis."
}
OUTPUT:
["may", "or"]

INPUT:
{
  "entity": "atelectasis",
  "sentence": "overall, however, there is a more
       focal airspace opacity in the left mid and
       lower lung, which may reflect asymmetric
       pulmonary edema or an infectious process,
       less likely atelectasis."
}
OUTPUT:
["less likely"]

... <redacted for brevity>

INPUT:
{
  "entity": "pleural effusion",
  "sentence": "a left pleural effusion and
       atelectasis obscure the left cardiac and
       hemidiaphragmatic contours more than the
       prior day."
}
OUTPUT:
[]

Examine the entity and sentence pair below. If the
     sentence also talks about other entities, first
      identify the part of the sentence that is
     talking about the query entity, and then
     extract phrases expressing uncertainty
     specifically related to that entity.
Return your output as a list: ["<word or phrase 1>",
      "<word or phrase 2>", ...]. If there are no
     uncertainty clues related to the given entity,
     return an empty list.

INPUT:
{
  "entity": {entity}
  "sentence": {sentence}
}
OUTPUT:
```

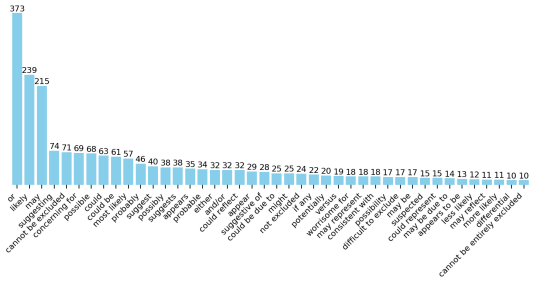


Figure A1: Our vocabulary of 42 common hedging phrases, including how many times each phrase was extracted for tentative finding-sentence pairs in LUNGUAGE.

We manually reviewed and corrected all cases where more than one phrase was extracted (568 finding-sentence pairs), since these were prone to mistakes. Furthermore, we noticed that the LLM occasionally extracted phrases which did not relate to the uncertainty of the finding, but rather to other relational attributes which were already extracted in LUNGUAGE: *onset*, *measurement* and *severity*. Phrases that matched these attributes for the same finding-sentence pair were filtered out. The phrase "difficult to assess" was extracted in 11 cases, but was removed since it relates more to visual limitations rather than uncertainty.

Figure A1 shows the number of times each of the 42 most common hedging phrases (i.e., occurring more than 10 times) were extracted from LUNGUAGE.

A.2. Reference ranking of hedging phrases

TrueSkill details The TrueSkill ranking system maintains a Bayesian belief in every item's skill by estimating two parameters: the average skill (μ) and the degree of confidence in the skill (σ), together characterizing a Gaussian distribution $\mathcal{N}(\mu, \sigma)$. Initially, we set $\mu = 25$ and $\sigma = \frac{25}{3}^2$, which is the default. TrueSkill additionally defines

a parameter β^2, which controls the variance of the item's performance around its average skill in a single match, set to $25/6$ per default.

When we use an LLM to compare two hedging phrases, we are essentially playing a match where one hedging phrase is the winner. After each match, TrueSkill updates both μ and σ using Bayesian inference: if an item performs better than expected, its μ increases, with the magnitude of this change depending on σ, and vice versa if it performs worse than expected. This mechanism is formalized in TrueSkill's update equations (Herbrich et al., 2006), which ensure that σ naturally shrinks as the system becomes more confident in its estimate, stabilizing each item's rating. Since the winner of a match is picked as the phrase that is closest to "finding is certainly present", a higher μ will be obtained for phrases closer to that part of the likelihood spectrum. In our case, we input all pairwise comparisons (across 861 pairs of hedging phrases, each repeated 10 times, with 4 LLMs) one after the other, in a random order. Since the final ranking is order-dependent, we repeat this 10 times with different random orders, averaging the final μ and σ across runs.

Prompting details The full prompt that was used to perform comparisons between pairs of hedging phrases is shown in Listing A2. We executed this prompt for each pair of hedging phrases, repeating 10 times with different example sentences (*sentence_1* and *sentence_2*), and with each of the four LLMs: *gemini-2.5-flash*, *gpt-4o*, *claude-sonnet-4*, *medgemma-27b-text-it*. For each LLM, we set the maximum number of tokens to 100, *temperature* to 1, *top_p* to 1. This time, we assigned zero thinking budget to Gemini, to pull all the models to the same level of reasoning capabilities.

Listing A2: Full prompt for hedging phrase comparison

```
SYSTEM: You are a radiologist who is ranking
     sentences expressing uncertainty.

TASK: You will be given two sentences from radiology
      reports. Each sentence contains a placeholder
     "<finding>", which represents a medical
     observation (e.g., consolidation, effusion,
     nodule). Each sentence includes a phrase that
     expresses the degree of certainty about the
     presence or absence of the finding.

Assume there is a certainty spectrum ranging from:
    "<finding> is certainly absent"
    to
    "<finding> is certainly present"
Your task is to identify which sentence is **closer
     to "<finding> is certainly present"** on this
     scale, using the context of the sentence. In
     other words, your task is to identify which
     sentence expresses a higher degree of certainty
      that the finding is present.

Respond with **only** the chosen sentence (
     sentence_1 or sentence_2).

Here are some examples.

---

Example 1:
INPUT:
{
  "sentence_1": "interstitial markings are prominent
       , suggest possible mild <finding>.",
  "sentence_2": "allowing for low inspiratory
       volumes, the <finding> is probably unchanged
       ."
}
OUTPUT:
"sentence_2"

---

Example 2:
INPUT:
{
  "sentence_1": "given the clinical presentation, <
       finding> must be suspected.",
  "sentence_2": "although this could represent
       severe <finding>, the possibility of
       supervening pneumonia or even developing ards
        must be considered."
}
OUTPUT:
"sentence_1"

---

Example 3:
INPUT:
{
  "sentence_1": "this could be either pneumonia in
       the left upper lobe or fissural <finding>.",
  "sentence_2": "the presence of a minimal left <
       finding> cannot be excluded, given blunting
       of the left costophrenic sinus.",
}
OUTPUT:
"sentence_1"

---

INPUT:
{
  "sentence_1": {sentence_1},
  "sentence_2": {sentence_2}
}

Which of the two sentences ("sentence_1" or "
     sentence_2") indicates that <finding> is more
     certainly present? Respond with your choice **
     only**.

OUTPUT:
```

Reference ranking To assess the robustness of our reference ranking (shown in Figure 3), we constructed an alternative ranking using only five comparisons per phrase pair. Compared to the original ranking in Figure 3, this reduced version showed an average absolute difference of 0.36 in μ and 0.76 in rank, with a Spearman correlation of 0.996 between both rankings. These results indicate that ten comparisons per pair provide a sufficiently stable and reliable ranking. We also compute pairwise agreement between the four LLMs across all 8610 comparisons, which is shown in Table A1. The average agreement scores are 0.865 for Gemini, 0.861 for GPT-4o, 0.866 for Claude, and 0.866 for MedGemma. Fleiss' Kappa between all four models is 0.860 and Krippendorff's Alpha is 0.722. These values indicate strong but imperfect agreement across models, highlighting the need

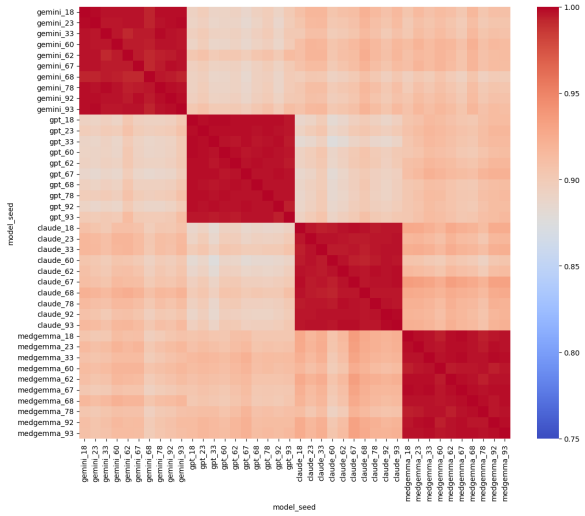


Figure A2: Pairwise Spearman rank correlations between TrueSkill rankings obtained for individual LLMs, across 10 seeds.

to integrate the judgments of all four LLMs when constructing the final ranking, as no single model can be assumed to provide the correct outcome in every case.

Table A1: Pairwise agreement between LLMs across 8610 hedging phrase comparisons.

Model	Gemini	GPT-4o	Claude	MedGemma
Gemini	1.000	0.861	0.868	0.866
GPT-4o	0.861	1.000	0.861	0.862
Claude	0.868	0.861	1.000	0.870
MedGemma	0.866	0.862	0.870	1.000

To further explore the variation across LLMs, we apply the TrueSkill algorithm to each individual LLM's set of comparisons (8610 each), repeating 10 times for different random orderings of the matches. We then compute pairwise Spearman rank correlations between the obtained rankings. The full correlation plot is show in Figure A2. Note that correlations are high for within-LLM comparisons, indicating that averaging across 10 ordering seeds should result in a stable ranking. Correlations are lower for across-LLM comparisons, once again indicating that relying on a single LLM to build our reference ranking would not be sufficient. As stated before, our final reference ranking (shown in Figure 3) is obtained by applying the TrueSkill algorithm to the full set of comparisons, repeating 10 times across ordering seeds and averaging the final μ and σ across runs. Figure A3 shows the Spearman rank correlation between each individual LLM's ranking across 10 seeds and the final reference ranking. Note that these correlations are generally higher than the pairwise correlations between individual LLMs from Figure A2.

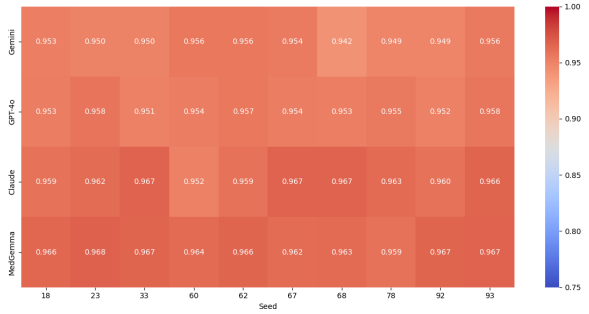


Figure A3: Spearman rank correlations with the reference ranking for TrueSkill rankings obtained for individual LLMs, across 10 seeds.

Expert evaluation Each participant was presented with 50 pairs of hedging phrases, with five example sentences per phrase. They were asked to choose the phrase that expresses a higher degree of certainty that the finding is present; an example is shown in Listing A3. The full survey, including detailed instructions for the evaluators, can be found on our Github repository, in the file `survey_uncertainty.pdf`. Table A2 shows the occupation and experience of the six experts involved in our expert evaluation study. Figure A4 shows the pairwise agreement between experts, and between experts and the reference ranking. Agreement is defined as the proportion of phrase pairs (out of 50) for which the expert's relative ordering matches that of the other expert (or of the reference ranking).

Listing A3: Example question in expert evaluation study

```
Phrase 1: difficult to exclude.
Example sentences:
- bilateral hilar vascular prominence is re-
    demonstrated with subtle <finding> in the left
    upper lung likely representing confluence of
    vasculature though a true nodule **difficult to
     exclude**.
- there is slight blunting of both costophrenic
    angles, felt most likely be due to overlying
    soft tissues, but a trace <finding> be **
    difficult to exclude**.
- no large <finding> is seen, although trace
    effusions are **difficult to exclude**.
- no large <finding> however, trace bilateral <
    finding>s **difficult to exclude**.
- mild <finding> is **difficult to exclude** in the
    correct clinical setting.

Phrase 2: appear.
Example sentences:
- the <finding> **appear** clear.
- the <finding> **appear** well inflated.
- the mediastinal and <finding> **appear** unchanged
    , allowing for differences in technique.
- mid <finding> **appear** intact.
- <finding> **appear** grossly intact.
```

A.3. Fitting each finding-sentence pair into the reference ranking

Algorithm Using Algorithm A1, we can fit any tentative finding-sentence pair into the reference ranking. Here, the draw probability between the target sentence t_{tar} with $(\mu_{tar}, \sigma_{tar})$, and an opponent phrase opp with $(\mu_{opp}, \sigma_{opp})$ is calculated as follows:

Type	Occupation	Experience
Writer	Radiologist	5 to 10 years
Writer	Radiologist	0 to 5 years
Reader	Internist	30 to 40 years
Reader	Oncologist	5 to 10 years
Reader	GP	30 to 40 years
Reader	GP	0 to 5 years

Table A2: Expert occupation and experience

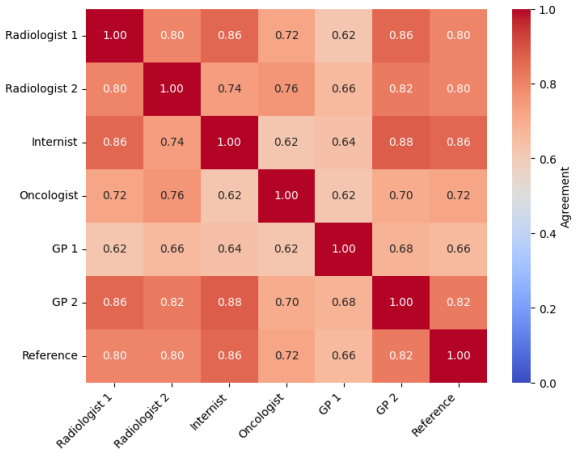


Figure A4: Pairwise agreement between experts and reference ranking, across 50 hedging phrase pairs.

$DrawProb(t_{tar}, opp) = exp(\frac{-(\mu_{tar}-\mu_{opp})^2}{2c^2})\sqrt{d}$, with $c^2 = 2\beta^2 + \sigma_{tar}^2 + \sigma_{opp}^2$, $d = 2\frac{\beta^2}{c^2}$, and $\beta^2 = 25/6$ (Herbrich et al., 2006).

Validation To validate our algorithm and the use of draw probability as an opponent selection strategy, we designed the following experiment. While the true rank of individual sentences in our dataset is unknown, we can assess performance using the hedging phrases in our reference vocabulary, whose ranks are known. Specifically, we simulate a modified reference ranking in which one of the 42 hedging phrases is temporarily excluded, then fit that phrase back into the ranking using a modified version of Algorithm A1. We consider two opponent selection strategies: (i) the *draw probability* strategy described previously, and (ii) a *random* strategy, where an opponent is chosen uniformly at random at each step. We also test multiple LLM configurations: one using *all* LLMs for the first K iterations (as in Algorithm A1) and then picking an LLM randomly, and another using a single LLM (*Gemini*, *GPT-4o*, *Claude*, or *MedGemma*) for all comparisons. Each configuration is repeated with 10 random seeds, which control sentence sampling, LLM selection, and opponent choice in the random setting. At each iteration, we compute the absolute distance between the current estimated rank of the phrase and its true rank in the reference ranking. If

Algorithm A1 Determine TrueSkill score μ for finding-sentence pair (f_{tar}, t_{tar}) based on reference ranking

Input: Target sentence t_{tar}, reference ranking $\mathcal{R}$, LLMs, K, N, max_steps $= 100$, patience $= 10$
Output: TrueSkill score $(\mu_{tar}, \sigma_{tar})$ for t_{tar}
Initialize $\mu_{tar} \leftarrow 25$, $\sigma_{tar} \leftarrow \frac{25}{3}^2$, $step \leftarrow 0$, $s \leftarrow 0$
Initialize opp_counts[phrase] $\leftarrow 0$, $\forall$ phrase $\in \mathcal{R}$
while $step <$ max_steps **and** $s <$ patience **do**
  Select $opp \in \mathcal{R}$ with $max\ DrawProb(t_{tar}, opp)$
  if $opp\_counts[opp] \geq N$ **then** skip
  $opp\_counts[opp] \leftarrow opp\_counts[opp] + 1$
  Randomly select sentence t_{opp} containing opp
  if $step < K$ **then**
    $models \leftarrow$ all LLMs
  else
    $models \leftarrow$ randomly pick 1 LLM
  end if
  Compare t_{tar} vs t_{opp} using $models$
  Update TrueSkill $(\mu_{tar}, \sigma_{tar})$ for t_{tar}
  Recalculate rank r_{tar} of μ_{tar} among $\mathcal{R}$
  If r_{tar} unchanged **then** $s \leftarrow s + 1$ **else** $s \leftarrow 0$
  $step \leftarrow step + 1$
end while

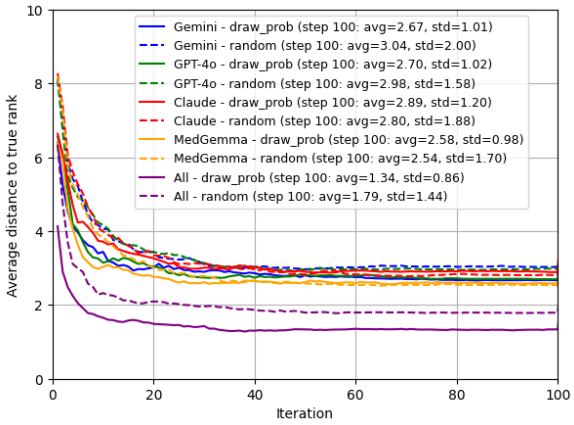
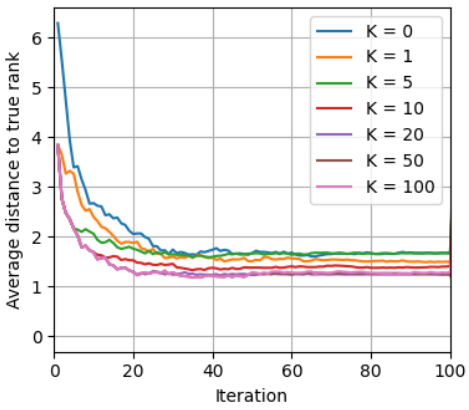


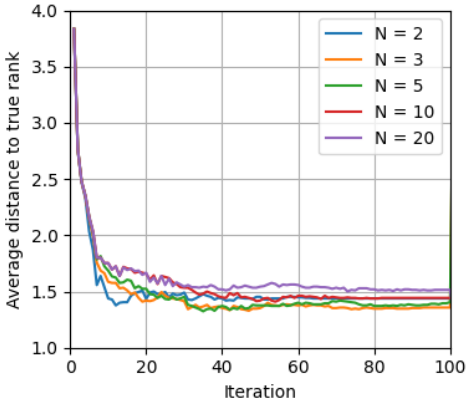
Figure A5: Average distance to the true rank across iterations of variants of Algorithm A1. We test different variants of the opponent selection strategy–*draw probability* vs. *random*–and LLMs used for comparisons–*all* or *single-LLM* (Gemini, GPT-4o, Claude, or MedGemma).

the algorithm converges before 100 iterations, the final rank is used for all subsequent iterations. We then average the absolute rank distances across all 42 phrases and 10 seeds. In these experiments, K was set to 10 and N to 5.

Figure A5 presents the resulting performance across all configurations. Note that the strategy which uses *draw probability* and *all* LLMs performs best, which is indeed the strategy implemented in Algorithm A1. While the *random* strategy performs similarly to *draw probability* in the *single-LLM* setting, it leads to less stable runs, as evidenced by the standard deviation of the distance to the true rank at step 100, averaged across all phrases, which is shown in Figure A5 as well.



(a) K



(b) N

Figure A6: Hyperparameter tuning results for K and N.

Hyperparameters Algorithm A1 contains hyperparameter K, which controls the number of iterations for which all four LLMs perform the phrase-sentence comparison, and hyperparameter N, which decides the number of times each opponent can be selected for comparison. Following the same experiment setup outlined above, with the *draw probability* strategy and including *all* LLMs, we perform hyperparameter optimization across the hedging phrases in our vocabulary. For K, we ran Algorithm A1 for $K \in [0, 1, 5, 10, 20, 50, 100]$ with $N = 5$. Figure A6a shows the average distance to the true rank in the reference ranking, averaged across all phrases. Setting $K = 10$ forms the right balance between cost-efficiency and performance. We performed the same experiment for $N \in [2, 3, 5, 10, 20]$, with $K = 10$, ultimately choosing $N = 5$ (Figure A6b).

Results and cost Applying Algorithm A1 to the 2,066 tentative finding-sentence pairs in LUNGUAGE incurred a total cost of $92.16, averaging $0.045/pair. Specifically, we spent $5.59 on Gemini, $39.49 on GPT-4o, and $45.12 on Claude, while MedGemma incurred no cost as it was run locally. On average, 24.61 ranking steps were required per finding-sentence pair, ranging from 10 to 100. The average μ value is 22.62, with a minimum of 0.41 and a maximum of 48.58, compared to a minimum of 7.07 and a maximum of 43.44 in the reference ranking (see Figure 3). The average rank when fit into the reference ranking is 23.87, with a minimum rank of 1 and a maximum rank of 43.

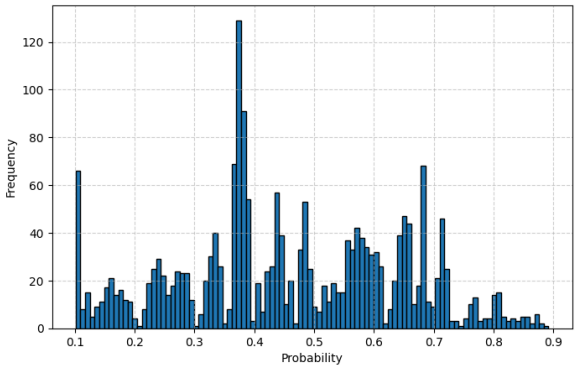


Figure A7: Histogram of probabilities in the tentative portion of LUNGUAGE^{++}.

A.4. Map to probability scale

We use an expert-informed sigmoid mapping (described in Section 3.4) to transform each μ into its probability value p. A histogram of the probabilities in the tentative portion of LUNGUAGE^{++} is shown in Figure A7.

This mapping is anchored by the expert-defined probabilities for the phrases *less likely* and *most likely*. Two radiologists received the following instruction:

```
On a scale of certainly absent (0) to certainly
    present (100), where would you place the phrase
     <phrase>, as it relates to the <finding> in
    each of the following sentences? Keep in mind
    that your answer may differ based on the
    context of the sentence.
```

They were asked to assign such probabilities for 10 example sentences, for both phrases. An example sentence for *less likely* includes: “Minimal blunting of the right costophrenic sulcus is more suggestive of similar slight atelectatic change, less likely persistent trace <finding>”. An example sentence for *most likely* includes: “Streaky predominantly right-sided mid and lower lung opacities are seen, most likely due to <finding>”. The full instructions are included in the survey, which can be found on our Github repository, in the file `survey_uncertainty.pdf`.

B. Implicit Uncertainty

B.1. Diagnostic Pathways

We present the diagnostic pathways defined in *Pathway Expansion Framework*. In practice, the pathways encode context primarily along two axes: (i) **imaging view and patient positioning**, which determine which radiographic signs can be expected or meaningfully interpreted (e.g., pleural effusion on an erect PA view typically presents with costophrenic angle blunting, whereas on a supine AP view it more often appears as diffuse haziness over the pleural space); and (ii) **clinical acuity together with morphologic/anatomic subtype**, which jointly modulate how a diagnosis presents (e.g., pneumothorax shows peripheral pleural air with absent pulmonary lung markings, whereas tension pneumothorax additionally shows mediastinal shift; pleural effusion may be non-loculated or loculated). These conditioning factors are encoded in the pathway dictionary. During expansion, the *Pathway Expansion Framework* uses these pathways to recover omitted intermediate findings in a report. Table B3 and Figures B1 to B2 present representative pathway variants for the 14 diagnoses. In the figures, each panel depicts the diagnosis as the root and its subfindings as children, with arrows indicating expert-defined dependencies. For example, in congestive heart failure the pathway links the root to cardiomegaly, pulmonary edema, and dyspnea (congestive heart failure $\rightarrow$ cardiomegaly + pulmonary edema + dyspnea). A dictionary with all pathway variants is also released on our Github repository, in the file `dx_pathway.csv`.

B.2. Cascading Expansion Framework

B.2.1. Blacklist for Finding Deduplication

We merge near-duplicate findings into a single canonical form using cosine similarity from a clinical embedding model (e.g., BioLord (Remy et al., 2023)) with a threshold of 0.9. Even with this high threshold, purely lexical similarity can still spuriously merge clinically distinct statements that differ along critical axes (e.g., ***left*** *lower lobe new consolidation* vs. ***right*** *lower lobe new consolidation*). To avoid false merges of distinct concepts, we apply an explicit blacklist with two layers: (i) exact-pair exclusions and (ii) pattern-level rules. We use a manually curated blacklist of mutually incompatible pairs, which is invoked when two candidates belong to the same study and to the same coarse anatomic region but differ along key discriminative axes such as laterality, lobar/segmental location, diagnostic status, or pathophysiologic mechanism. A candidate pair is merged only if it passes the similarity threshold and does not match any blacklist entry; otherwise both findings are preserved. Representative blacklist entries and patterns are summarized in Table B1.

B.2.2. Details of Pathway Matching

Algorithm B1 describes the pathway matching procedure, which aligns each input tuple $(\textit{finding}, \textit{loc}, \textit{attr}, \textit{view})$ to at most one pathway variant from Table B3 and Figures B1 and B2. The full code can be found on our Github repository.

For each row from LUNGUAGE, we first apply `NormalizeInput` to obtain a normalized tuple (f', ℓ', a', v'), mapping raw report phrases to the LUNGUAGE vocabulary (e.g., "RUL" $\rightarrow$ "right upper lung," "TB" $\rightarrow$ "tuberculosis") and mapping view metadata to a standardized `view_information` string (AP, PA, LATERAL, erect, supine). In parallel, we apply `NormalizeDict` to the pathway dictionary $\mathcal{D}$, yielding $\mathcal{D}'$ with normalized findings, locations, attributes, and views for every pathway entry.

Given (f', ℓ', a', v') and $\mathcal{D}'$, the `BestPathwayMatch` routine proceeds in three compatibility steps. First, it performs a *finding-level lexical match*: we restrict candidates to those pathways whose normalized finding term is identical to f'. This ensures that, for example, tuples with finding "fracture" only compete among fracture-related pathways, and tuples with finding "pleural effusion" only among effusion pathways.

Second, we enforce *location and attribute compatibility*. For example, location compatibility distinguishes skeletal fractures from device fractures even when the normalized finding f' is "fracture". Attribute terms then select more specific variants: "loculated" triggers the loculated-effusion pathway, "tension" the tension-pneumothorax pathway, "compression" the compression-fracture pathway, and acuity modifiers such as "acute," "old," or "healed" select the corresponding fracture branch.

Finally, we check *view compatibility* using `view_information`, retaining only variants that are observable under the given view (e.g., erect versus supine presentations of pleural effusion, as discussed earlier). By construction, these compatibility rules are defined so that at most one pathway variant remains for any normalized input tuple. If a unique compatible variant exists, `BestPathwayMatch` returns it as $c^\star$; otherwise it returns *none*.

B.2.3. Status Conflict Resolution

During pathway expansion we observed status conflicts for the same finding, and we also identified inconsistencies within original reports. In Table B2, we categorize conflicts by *source* and *type* and count their occurrence in LUNGUAGE. Sources are `original_vs_expansion` (an expanded node contradicts an explicitly stated node),

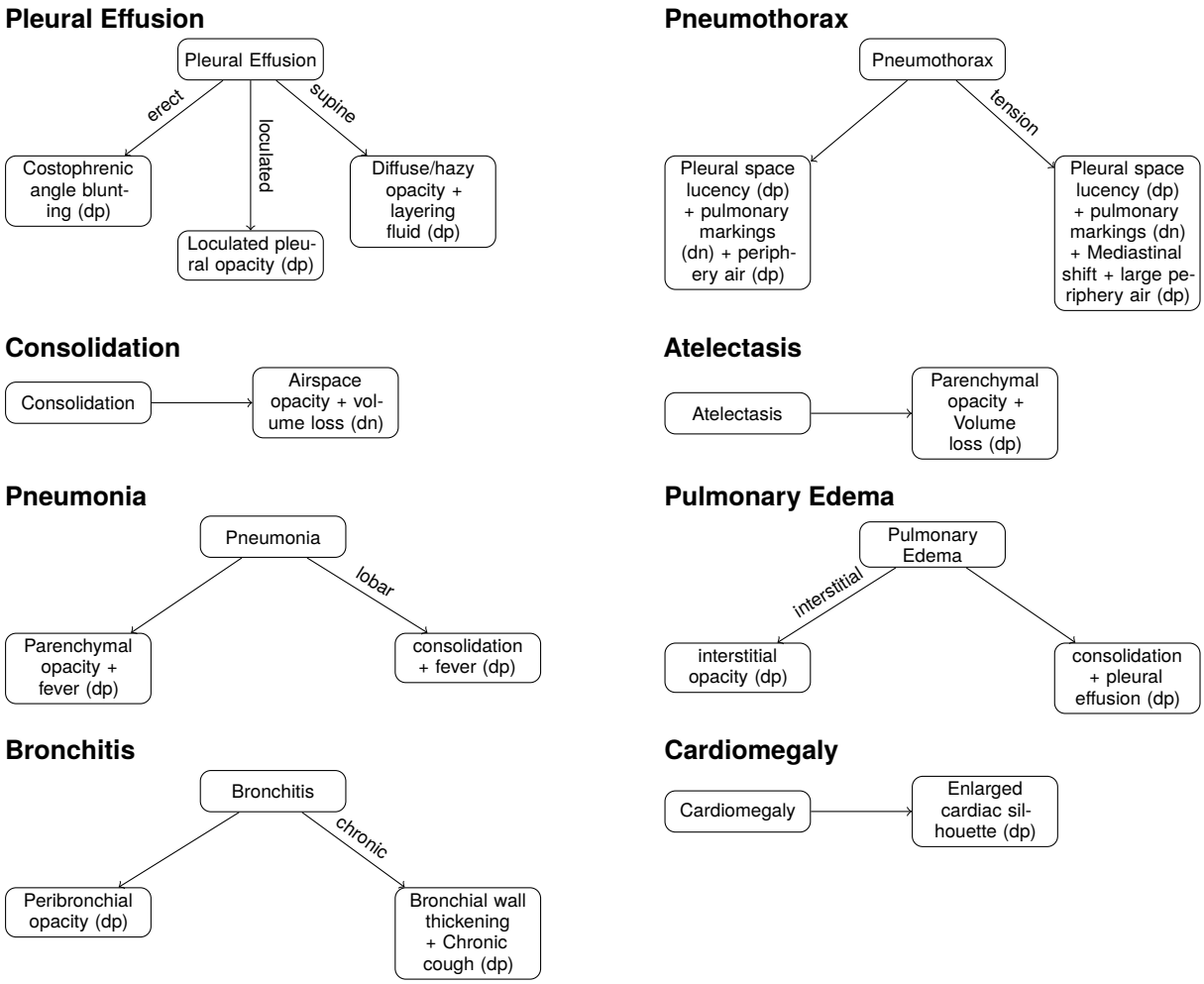


Figure B1: Diagnostic pathway panels (Part 1 of 2) illustrating exemplar pathways from the hand-crafted dictionary used by the pathway matching algorithm. Each panel is a small directed graph whose root node is a normalized diagnosis (*finding*) and whose children are normalized sub-finding patterns over location, attributes, view, and (optionally) clinical context (e.g., patient symptom). Edges are annotated with the attribute or view modifiers that the matcher uses when checking joint compatibility between an input tuple (*finding*, *loc*, *attr*, *view*) and a pathway variant. Here, *dp* stands for definitive positive, while *dn* stands for definitive negative. Shown pathways correspond to: Pleural Effusion, Pneumothorax, Consolidation, Atelectasis, Pneumonia, Pulmonary Edema, Bronchitis, and Cardiomegaly. For Pleural Effusion, the branches encode erect versus supine presentations and a loculated variant. For Pneumothorax, the variants capture simple pneumothorax, tension pneumothorax, and hydropneumothorax. For Pneumonia and Pulmonary Edema, the pathways encode typical parenchymal and interstitial/alveolar patterns (consolidation) together with key clinical or co-occurring radiographic findings (e.g., fever, pleural effusion). Bronchitis is represented as an airway-centered process, while Consolidation, Atelectasis, and Cardiomegaly map directly to their canonical imaging signatures. All pathway structure shown here is taken directly from the released pathway dictionary and is the same structure searched by `BestPathwayMatch`.

`original_vs_original` (two original nodes disagree), and `expansion_vs_expansion` (two expanded nodes disagree). Types are `polarity` (dp vs. dn), `certainty_positive` (dp vs. tp), `certainty_negative` (dn vs. tn), `duplicate_pos` (tn vs. tn) and `duplicate_neg` (tp vs. tp).

Conflict detection operates at the granularity of (`entity`,`location`) within each report with the same `study_id`. Across the expanded LUNGUAGE dataset (19,216 rows), expansion-related conflicts were rare overall, totaling 616 cases (**3.2%**): 165 (**0.9%**) `original_vs_expansion` and 451 (**2.3%**) `expansion_vs_expansion`. Within `original_vs_expansion` (n=165), conflicts due to polarity of definitive findings dominated, with `polarity` accounting for 107 (64.8%) conflicts, `certainty_pos` for 34 (20.6%), and `certainty_neg` for 24 (14.5%). Within `expansion_vs_expansion` (n=451), conflicts were dominated by duplicates of positively tentative findings: `polarity` accounted for 103 (22.8%) conflicts, `certainty_pos` for 133 (29.5%), `certainty_neg` for 16 (3.5%), `duplicate_pos` for 181 (40.1%), and `duplicate_neg` for 18 (4.0%). In addition, some inconsistencies were already present in the original reports: `orig-`

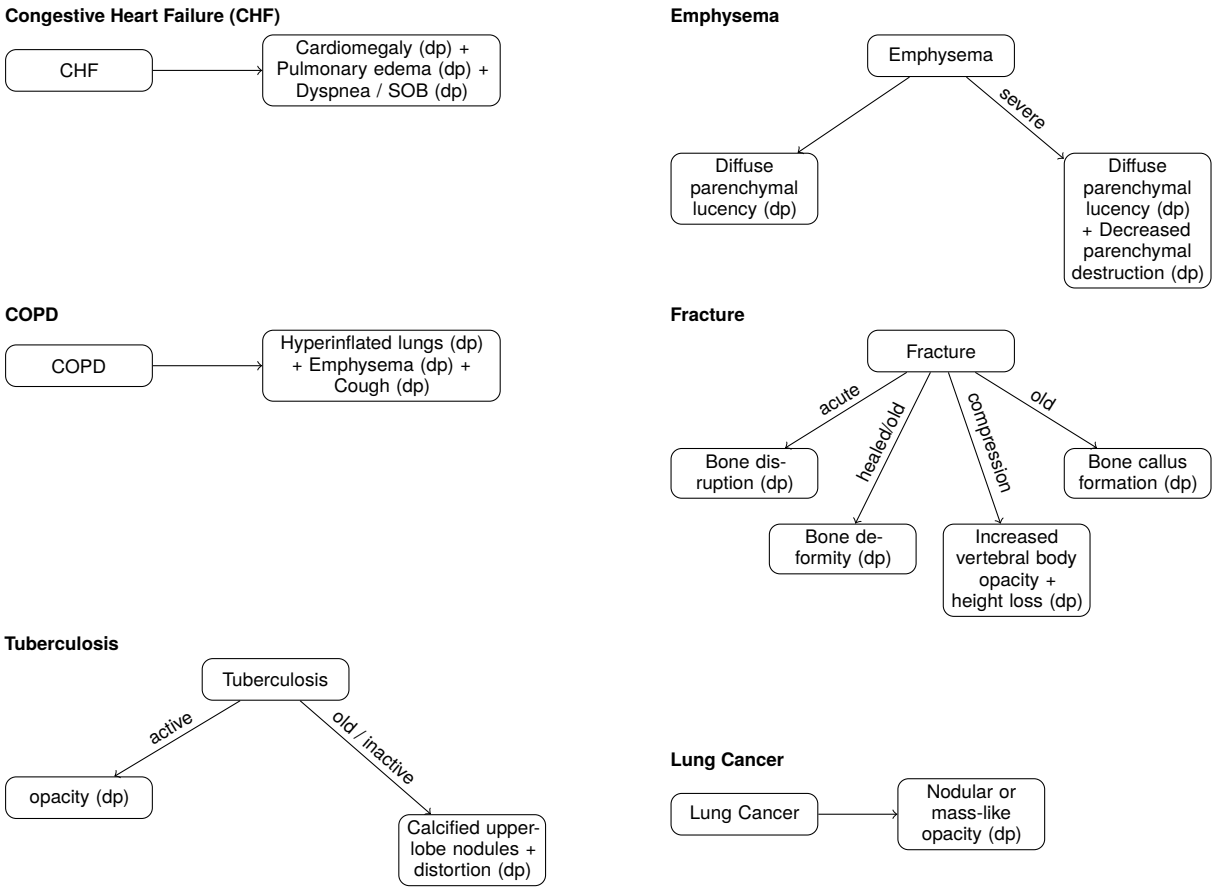


Figure B2: Diagnostic pathway panels (Part 2 of 2) illustrating exemplar pathways from the hand-crafted dictionary used by the pathway matching algorithm. Shown pathways correspond to CHF, Emphysema, COPD, Fracture, Tuberculosis, and Lung Cancer. CHF is modeled as a cardiopulmonary congestion cluster, where cardiomegaly, pulmonary edema, and dyspnea co-occur during exacerbation. Emphysema encodes a severity axis from diffuse parenchymal lucency to severe parenchymal destruction with decreased peripheral vascularity. COPD is represented as a chronic obstructive cluster combining hyperinflated lungs, emphysematous change, and chronic cough. Fracture pathways separate acute cortical breaks, old or healed deformity, and vertebral compression patterns. Tuberculosis includes an explicit disease-state axis, contrasting active parenchymal opacity with old or inactive upper-lobe scarring that shows calcified nodules and architectural distortion. Lung Cancer is represented by a focal nodular or mass-like opacity that may reflect primary or metastatic disease.

Algorithm B1 Pathway Matching

Input: (f, ℓ, a, v) from LUNGUAGE, pathway dict $\mathcal{D}$ *// f: finding, ℓ: location, a: attributes, v: view*
Output: $c^\star$ or *none*

function SELECTVARIANT($f, \ell, a, v, \mathcal{D}$)
  $(f', \ell', a', v') \leftarrow \mathrm{NormalizeInput}(f, \ell, a, v)$ *// normalize row-level finding, location, attributes, and view*
  $\mathcal{D}' \leftarrow \mathrm{NormalizeDict}(\mathcal{D})$ *// normalize findings, locations, attributes, and views in all pathways*
  $c^\star \leftarrow \mathrm{BestPathwayMatch}(\mathcal{D}', f', \ell', a', v')$
  return $c^\star$
end function
function BESTPATHWAYMATCH($\mathcal{D}', f', \ell', a', v'$)
  $C \leftarrow \{\, c \in \mathcal{D}' : \mathrm{Find}(c) = f' \,\}$ *// finding-level lexical match*
  if $C = \emptyset$ **then**
    return *none*
  end if
  $C \leftarrow \{\, c \in C : \mathrm{LocOK}(c, \ell') \wedge \mathrm{AttrOK}(c, a') \wedge \mathrm{ViewOK}(c, v') \,\}$ *// joint compatibility on location, attributes, and view*
  if $|C| = 1$ **then**
    return the sole element of C
  else
    return *none* *// ambiguous or incompatible (should not occur since pathways are independent)*
  end if
end function

`inal_vs_original` totaled 158 cases (**0.8%**). These conflicts were dominated by positive certainty conflicts: `polarity` accounted for 10 (6.3%) conflicts, `certainty_pos` for 74 (46.8%), `certainty_neg` for 2 (1.3%), `duplicate_pos` for 70 (44.3%), and `duplicate_neg` for 2 (1.3%). Overall conflicts across all sources summed to 774 (**4.0%** of the initial expanded LUNGUAGE dataset).

Resolution follows a deterministic policy: (1) if a group contains both original and expansion rows, retain the originals and discard the expansions (treat the report text as the clinical source of truth); (2) if the remaining rows show a pure polarity clash with only `dp` and `dn`, drop the group; (3) otherwise select the row with the highest `prob`. When multiple rows tie on `prob`, break ties by a status priority that reflects clinical certainty, `dp` > `tp` > `tn` > `dn`. After resolution, **18,810** rows remained (**97.9%** of 19,216), implying **406** removals (**2.1%**); all remaining inconsistencies were eliminated, preserving log-

Category	Non-merge pairs (substring-level)
Laterality and anatomical direction	(left, right), (left-sided, right-sided), (left-sided, right), (left, right-sided), (upper, lower), (mid, lower), (upper, mid), (upper, middle), (middle, lower), (anterior, posterior), (anterior, lateral), (posterior, lateral), (superior, inferior), (apical, basal), (central, peripheral), (proximal, distal), (medial, lateral), (ventral, dorsal)
Cardiopulmonary contextual conflicts	(cardio, pulmonary), (mediastinal, pleural), (pericardial, pleural)
Disease mechanism divergence	(effusion, pneumothorax), (effusion, edema), (effusion, atelectasis), (effusion, consolidation), (effusion, pneumonia), (consolidation, opacification), (consolidation, atelectasis), (consolidation, edema), (atelectasis, pneumonia), (atelectasis, aeration), (atelectasis, opacity), (pneumonia, edema)
Mass vs. airspace process	(mass, consolidation), (mass, opacity), (mass, atelectasis)

Table B1: Manually curated blacklist of substring pairs that prevent merges during high-precision deduplication.

ical consistency and clinical validity. Algorithm B2 details the procedure.

Conflict	Type	exp ↔ exp	ori ↔ exp	ori ↔ ori
dn ↔ tn	`certainty_neg`	16	0	2
dp ↔ tp	`certainty_pos`	133	34	74
tn ↔ tn	`duplicate_neg`	18	0	2
tp ↔ tp	`duplicate_pos`	181	24	70
dp ↔ dn	`polarity`	103	107	10
		451	165	158

Table B2: Pathway expansion conflicts by source and type. *ori* and *exp* denote original and expansion.

Algorithm B2 Conflict Resolution

Input: Expanded dataset D with columns: `study_id`, `entity`, `location`, `status` $\in$ {`dp`, `tp`, `tn`, `dn`}, `prob`, `source` $\in$ {`original`, `expansion`}
Output: `resolved_df`

keys ← [`study_id`, `entity`, `location`]
`RES` ← [ ]
for all group $G \subset D$ by `keys` **do**
  if $|G| < 2$ **then continue**
  end if
  `orig` ← { $r \in G \mid r.$`source` $=$ `original` }
  `expd` ← { $r \in G \mid r.$`source` $=$ `expansion` }
  — Case 1: original vs. expansion —
  if `orig` $\neq \varnothing$ **and** `expd` $\neq \varnothing$ **then**
    Rule1: keep originals only; discard expansions
    append all rows in `orig` to `RES`
    continue
  end if
  — Case 2: only originals or only expansions —
  $R \leftarrow$ `orig` if `orig` $\neq \varnothing$ else `expd`
  $has\_dp \leftarrow (\exists r \in R : r.$`status` $=$ `dp`$)$
  $has\_dn \leftarrow (\exists r \in R : r.$`status` $=$ `dn`$)$
  $has\_tp \leftarrow (\exists r \in R : r.$`status` $=$ `tp`$)$
  $has\_tn \leftarrow (\exists r \in R : r.$`status` $=$ `tn`$)$
  Rule2: drop if pure polarity clash
  if has_dp **and** has_dn **and not**(has_tp **or** has_tn) **then**
    continue
  end if
  Rule 3: pick highest probability; dp > tp > tn > dn
  $p^* \leftarrow \max\{ r.$`prob` $: r \in R \}$
  if $\exists r \in R : r.$`prob` $= p^* \wedge r.$`status` $=$ `dp` **then**
    append any such r to `RES`
  else if $\exists r \in R : r.$`prob` $= p^* \wedge r.$`status` $=$ `tp` **then**
    append any such r to `RES`
  else if $\exists r \in R : r.$`prob` $= p^* \wedge r.$`status` $=$ `tn` **then**
    append any such r to `RES`
  else
    append any $r \in R$ with $r.$`prob` $= p^*$ to `RES`
  end if
end for
`resolved_df` ← Concat(`RES`)
return `resolved_df`

Diagnosis	Specific diagnosis	Diagnostic pathways
Pleural Effusion	Non-loculated pleural effusion *(one of nine pathways)*	`view: ap, pa, lateral > ent: blunting > status: dp > loc: costophrenic angle && ent: opacity > status: dp > loc: pleural space > attr: hazy, diffuse`
	Loculated pleural effusion	`view: ap, pa, lateral > ent: opacity > status: dp > loc: pleural space > attr: loculated`
Pneumothorax	Pneumothorax	`view: ap, pa, lateral > ent: lucency > status: dp > loc: pleural space && ent: marking > status: dn > loc: pulmonary && ent: air > status: dp > loc: lung periphery`
	Tension pneumothorax	`Pneumothorax && ent: shift > status: dp > loc: mediastinal && attr: large amount`
Consolidation		`view: ap, pa, lateral > ent: opacity > status: dp > loc: airspace && ent: volume loss > status: dn`
Atelectasis		`view: ap, pa, lateral > ent: opacity > status: dp && ent: volume loss > status: dp`
Pneumonia	Pneumonia	`view: ap, pa, lateral > ent: opacity > status: dp && ent: fever > status: dp`
	Aspiration pneumonia	`Pneumonia > loc: lung > attr: dependent portion`
	Lobar pneumonia	`view: ap, pa, lateral > ent: consolidation > status: dp && ent: fever > status: dp`
Pulmonary Edema	Interstitial pulmonary edema	`view: ap, pa, lateral > ent: opacity > status: dp > loc: interstitial`
	Alveolar pulmonary edema	`view: ap, pa, lateral > ent: consolidation > status: dp && ent: pleural effusion > status: dp`
Bronchitis	Bronchitis	`view: ap, pa, lateral > ent: opacity > status: dp > loc: peribronchial`
	Chronic Bronchitis	`view: ap, pa, lateral > ent: thickening > status: dp > loc: bronchial wall && ent: cough > status: dp`
Cardiomegaly		`view: ap, pa > ent: heart size > status: dp > loc: cardiothoracic > attr: increased`
CHF		`view: ap, pa, lateral > ent: cardiomegaly > status: dp && ent: pulmonary edema > status: dp && ent: dyspnea > status: dp`
Emphysema	Emphysema	`view: ap, pa, lateral > ent: lucency > status: dp > loc: lung parenchyma > attr: diffuse`
	Severe emphysema	`Emphysema && ent: pulmonary vascularity > status: dp > attr: decreased && ent: destruction > status: dp > loc: lung parenchyma`
COPD		`view: ap, pa, lateral > ent: hyperinflation > status: dp > loc: lungs && ent: emphysema > status: dp && ent: cough > status: dp`
Fracture	Acute fracture	`view: ap, pa, lateral > ent: disruption > status: dp > loc: bone`
	Chronic/Old fracture	`view: ap, pa, lateral > ent: callus formation > status: dp > loc: bone`
	Healed fracture	`view: ap, pa, lateral > ent: deformity > status: dp > loc: bone`
	Spinal/Compression fracture	`view: ap, pa, lateral > ent: opacity > status: dp > loc: vertebral body > attr: increased && ent: loss of height > status: dp`
Tuberculosis	Active tuberculosis	`view: ap, pa, lateral > ent: opacity > status: dp`
	Chronic/Non-active tuberculosis	`view: ap, pa, lateral > ent: nodules > status: dp > loc: bilateral upper lung > attr: calcified && ent: architectural distortion > status: dp > loc: bilateral upper lung`
Lung Cancer		`view: ap, pa, lateral > ent: opacity > status: dp > attr: nodular`

Table B3: Diagnostic pathways. `view` denotes projection or patient orientation; `ent`, `status`, `loc`, and `attr` indicate the entity, status (`dp` is definitive positive, `dn` is definitive negative), anatomical location, and attributes (e.g., morphology, distribution, measurement). Pathways are ordered sequences joined by "`>`", with required co-occurrence marked by "`&&`". We display one representative pathway out of nine defined for pleural effusion and, in total, 33 pathways spanning 14 diagnoses.