Paint It, Black: Evaluating LLMs for De-Identification of Multimodal Clinical Documents

Francesco Zangrillo [*] **Alessio Cocchieri** [*] **Giacomo Frisoni** [*†]
Davide Freddi [*] **Luca Ragazzi** [*] **Alberto Aguzzi** ♫ ♪ **Diego Raimondo** ♫
Renato Seracchioli ♫ ♪ **Gianluca Moro** [*]
Department of Arts, University of Bologna, Italy
Department of Computer Science and Engineering, University of Bologna, Italy
♫ IRCCS Azienda Ospedaliero-Universitaria di Bologna, Italy
♪ Department of Medical and Surgical Sciences, University of Bologna, Italy

Abstract

Secondary use of clinical data requires removing Protected Health Information (PHI) under regulations such as HIPAA. While de-identification is well studied for text, clinical records are often structured PDFs, where PHI appears across complex layouts, tables, and headers. We introduce Paint-It-Black (PIB), the first benchmark for end-to-end, pixel-level PHI de-identification in clinical PDFs. PIB contains physician-validated synthetic documents with diverse layouts, document types, and controlled visual degradations, annotated with HIPAA-aligned labels and bounding boxes. We evaluate locally deployable Multimodal Large Language Models (MLLMs) and find that even the strongest remain below the reliability required for clinical use. A conventional OCR+NER baseline performs substantially worse than the best MLLMs, highlighting the inherently multimodal nature of clinical PHI de-identification.[1]

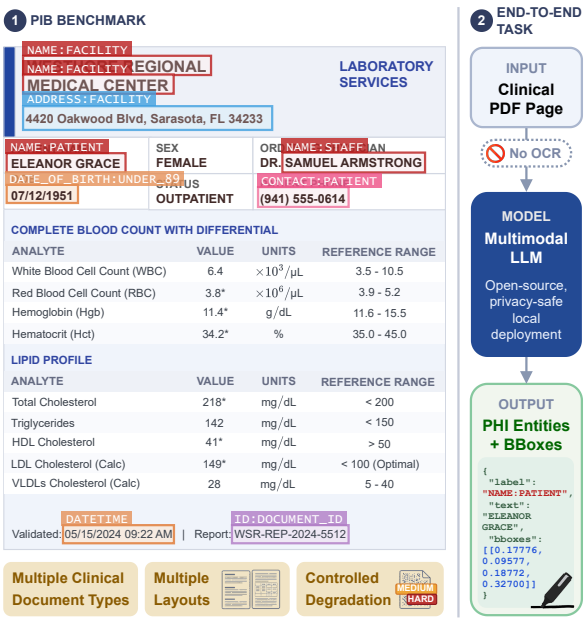


Figure 1: **PIB frames PHI de-identification as end-to-end multimodal NER.** Given a clinical PDF page image, a model must jointly predict HIPAA-aligned entity labels and normalized bounding-box coordinates.

1 Introduction

The clinical record is the atomic unit of modern medicine. Every diagnosis, prescription, and follow-up visit produces a document that encodes, together with medically irreplaceable information, a dense set of personal identifiers (patient names, dates of birth, facility addresses, etc.), collectively regulated under the umbrella of Protected Health Information (PHI). Under HIPAA in the United States[2] and Article 9 of the GDPR in the European Union (Vuković et al., 2022), health data is classified as a special category of personal data requiring elevated protection, and its secondary use for research or AI development is permissible only after proper de-identification or anonymization.

The NLP community recognized this necessity early. A long lineage of shared tasks and annotated resources has produced increasingly accurate text-based de-identification systems, progressing from rule-based dictionaries and conditional random fields to fine-tuned clinical transformers and, recently, generative Large Language Models (LLMs) that approach or exceed human recall (Aghakasiri et al., 2025). Yet this trajectory harbors a crucial assumption that is rarely interrogated: that clinical records are plain text. In practice, they are not.

The overwhelming majority of clinical documents that circulate in hospitals arrive as PDF files. Feeding such documents through an automatic de-identification system thus requires, implicitly or explicitly, a spatial reasoning step: the system must locate the PHI on the page before it can obfuscate or pseudonymize it. Current approaches address

[*] Equal contribution (co-first authorship)
[†] Corresponding author: giacomo.frisoni@unibo.it
[1] Code, data: https://disi-unibo-nlp.github.io/paint-it-black/ (MIT License)
[2] U.S. Department of Health and Human Services

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 12865–12902
October 24-29, 2026 ©2026 Association for Computational Linguistics

this by inserting an intermediate OCR stage, converting the document image back into text and then running a standard named entity recognition (NER) pipeline (Kocaman et al., 2023).

The emergence of Multimodal LLMs (MLLMs), that jointly process page images and text, opens a qualitatively different paradigm: end-to-end spatial PHI detection, in which the model receives a rendered document image and *directly* outputs both the entity label and its bounding-box coordinates on the page. This framing casts PHI de-identification as a form of visually grounded NER, demanding not only the semantic comprehension evaluated by text benchmarks but also the fine-grained spatial localization that existing evaluations do not test. Whether current MLLMs are actually capable of this joint operation—and how their capabilities degrade under the visual noise endemic to real hospital workflows—remains entirely uncharted.

To fill this gap, we present **Paint-It-Black (PIB)**, the first benchmark for end-to-end, spatially grounded PHI de-identification in clinical documents (see Fig. 1). At its core lies a corpus of synthetic PDFs generated through an LLM-based pipeline whose outputs were guided and filtered by practicing physicians to ensure clinical realism. Each document is annotated with bounding boxes and labels under a fine-grained HIPAA-aligned taxonomy, and paired with visually degraded counterparts simulating real-world acquisition conditions, enabling controlled evaluation across difficulty tiers. Building on this resource, we benchmark diverse open-source, locally deployable MLLMs—the only deployment paradigm compatible with hospital privacy constraints—under a rich metric suite. We also evaluate the conventional OCR+NER pipeline and find it trails every MLLM, showing that spatial PHI de-identification cannot be reduced to a text-extraction problem and instead demands genuine end-to-end multimodal reasoning.

2 Related Work

PHI Detection The dominant paradigm for PHI detection treats the problem as a text-span classification task, with a clear progression from domain-adapted encoders to instruction-tuned LLMs and multi-stage compliance frameworks (Singh et al., 2025) that achieve near-human recall on standard text benchmarks (Aghakasiri et al., 2025). A recent milestone in this fixed-taxonomy tradition is OpenAI Privacy Filter,[3] a 1.5B-parameter mixture-of-experts model that repurposes a generatively pre-trained language model as a bidirectional token classifier over eight Personal Identifiable Information (PII) categories—a superset of PHI. A parallel line of work shifts the detection bottleneck from supervised fine-tuning to open-vocabulary extraction, conditioning the model at inference time on natural-language descriptions of the target entity types rather than a fixed label set. GLiNER (Zaratiana et al., 2024, 2025) follows this paradigm using a bidirectional encoder. GLiNER2-PII (Zaratiana et al., 2026) specializes in a 42-type PII schema, outperforming Privacy Filter. In the medical domain, OpenBioNER-v2 (Cocchieri et al., 2026a) achieves competitive PHI detection with only 110M parameters by injecting entity type descriptions. Despite their diversity, all of the above operate exclusively on text strings, presupposing a reliable text extraction front-end. Spatial parsers such as LiteParse[4] partially address this by recovering bounding-box-aligned text spans directly from the PDF rendering engine, but resort to OCR for figures and scanned pages, reintroducing error compounding. Visual document retrievers such as ColPali/ColQwen (Faysse et al., 2025) sidestep OCR entirely. However, their coarse granularity and query-matching objective make them unsuitable for exhaustive entity localization. To our knowledge, PIB is the first benchmark to evaluate whether MLLMs can jointly spatially ground and classify PHI directly on raw clinical PDF page images, with no dependence on parsers, OCR, or retrieval models.

Clinical Document Datasets The most widely used resources for clinical NLP—MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023)—distribute free-text notes that have been anonymized prior to release: all PHI has already been stripped and replaced with placeholders, leaving no identifiable spans for a de-identification system to detect. Extensions of these corpora that distribute PDF files, such as MIMIC-III-Ext-VeriFact-BHC,[5] do not originate from clinical practice; they are renderings of annotation guidelines used for human review workflows. The closest existing resource to our work is the JohnSnowLabs PDF De-

[3] https://github.com/openai/privacy-filter
[4] https://github.com/run-llama/liteparse
[5] https://doi.org/10.13026/abat-g475

Dataset Metric		Value (%)
Total Benchmark Page Instances		414
Base Split (Clean 144 DPI)		138 (33.3%)
Medium Split (7 Artifacts)		138 (33.3%)
Hard Split (11 Artifacts + Markup)		138 (33.3%)
Unique Source PDF Documents		82
Multi-Page PDFs		42 (51.2%)
Clinical Domain Doc Types		4
Annotation Summary (Invariant per Split)		
Total PHI Annotations		2,579
Unique Clinical Entity Types (Labels)		23
Average Annotations per Page		18.69
Annotations with Multi-Bbox Spans		7.8%
Demographic Distribution (Invariant per Split)		
Country	US	41 (38.3%)
	Canada	39 (36.4%)
	UK	29 (27.1%)
	Australia	29 (27.1%)
Sex	Female	90 (65.2%)
	Male	48 (34.8%)
Annotation Density per Page (Invariant per Split)		
1–5 annotations		11 (8.0%)
6–15 annotations		45 (32.6%)
16+ annotations		82 (59.4%)

Table 1: **Statistics of the PIB benchmark.** We report total page instances across the three difficulty splits. Because annotations are held fixed during visual degradation, statistics are invariant across all splits.

identification Dataset,[6] a synthetic collection of 50 generic medical-style PDFs covering a single document category with minimal layout variation and no benchmarking results. PIB addresses all these limitations, providing a synthetic, physician-validated benchmark covering multiple document categories with rich layout diversity, controlled visual degradation tiers, and the first systematic open-model benchmark for end-to-end spatial PHI detection.

3 The Paint-It-Black Benchmark

We detail the taxonomy, generation, and orchestration of the PIB benchmark. As shown in Table 1, PIB comprises 414 multimodal page instances, featuring 2,579 labels per split, designed to stress-test models under realistic operational hazards.

3.1 Label Taxonomy

PIB's annotation schema is derived from the HIPAA Privacy Rule's 18-category Safe Harbor specification,[7] adapted to address the morphological and layout complexities of multimodal document extraction. Following the hierarchical *i2b2* format (`CATEGORY:SUBCATEGORY`) (Stubbs and Uzuner, 2015), the taxonomy covers 23 fine-grained entity types organized into 7 macro-categories. **(1) NAME** spans five subtypes: **Patient**, **Staff** (medical personnel) **Associate** (family members and third parties), **Facility** (clinics, hospitals), and **Department**. Annotation includes partial identifiers such as isolated surnames or initials to prevent record-linkage attacks (Sweeney, 2002). **(2) DATETIME** captures all non-birth temporal references, from administrative header timestamps to isolated years in narrative prose. **(3) DATE_OF_BIRTH** and **(4) AGE** are stratified into **UNDER_89** and **OVER_89** subtypes, as ages exceeding 89 require a distinct redaction regime under Safe Harbor rules due to their direct identifying power. **(5) ADDRESS** splits geographic identifiers from street level to country granularity into **Patient** and **Facility** targets. **(6) CONTACT** covers phone/fax numbers and emails for **Patient**, **Facility**, and **Associate**; department and facility numbers are merged because distinguishing them is an ill-defined problem in clinical headers lacking syntactic separators. Finally, **(7) ID** spans seven alphanumeric subtypes (**Patient**, **Admission**, **Exam**, **Staff**, **Specimen**, **Device**, and **Document**).

Conversely, non-identifying clinical metadata (e.g., exam types, diagnoses, or clinical findings) is deliberately left unannotated to preserve downstream utility for medical NLP tasks.

3.2 Dataset Generation

To circumvent the ethical constraints of redacting real-world medical records, we designed a scalable pipeline to generate high-fidelity synthetic documents and expose them to realistic visual decay.

The seed corpus comprises 82 unique clinical documents (138 distinct pages) distributed across four domains selected by our clinical co-authors to mirror standard hospital workflows: Laboratory (*lab_reports*), Computed Tomography (*ct_reports*), Magnetic Resonance Imaging (*mri_reports*), and Gynecological Consultations (*gyn_reports*), balancing single-page (48.8%) and multi-page samples (51.2%). Structural variance is enforced by executing a full Cartesian product across discrete parameter axes, including geographic location, patient age, sex, clinical framing, and visual formatting aesthetics to prevent template-overfitting.

[6] https://github.com/JohnSnowLabs/pdf-deid-dataset

[7] U.S. Department of Health and Human Services

Generative Pipeline Each document is synthesized through a deterministic, five-stage execution flow designed to yield high-fidelity visual layouts. First, we compile a structured natural-language instruction from the parameterized coordinates of the Cartesian product. Second, we submit this prompt to Gemini-3-Flash via a multi-turn chat interaction (see the prompt in §A.3). Third, we isolate the markdown-fenced HTML payload from the response, validating strict compliance with target element geometries. Fourth, we compile the raw HTML into a standardized A4 PDF via WeasyPrint,[8] enforcing programmatic page-break and overflow constraints. Finally, the rendered PDF pages are re-injected into the model's multimodal context to visually evaluate layout coherence, executing HTML patch corrections if overflow or visual clipping is detected. This vision-loop drastically reduces truncated tables and header spillage, preventing layout corruption that would invalidate spatial bounding-box ground truths. Detailed descriptions of the generation pipeline and parameter grid are provided in §A.1 and §A.2.

PHI Embedding & Realism To enforce genuine semantic reading comprehension and prevent models from exploiting spatial shortcuts—such as static header matching—PHI tokens are embedded both structurally (within tabular demographic blocks or institutional footnotes) and contextually (interwoven within narrative clinical prose, such as references to historical procedures at external facilities), reflecting the heterogeneous signals present in real-world data (Domeniconi et al., 2016). To counteract synthetic text regularities, the prompt dynamically enforces typographic perturbations—such as transposed characters and dictation false-starts—affecting 60% of the documents. These flaws are strictly restricted to non-PHI segments to preserve target entity syntax. Concurrently, the visual formatting style parameter randomizes rendering aesthetics from stark hospital terminal layouts to modern typeset institutional letters, enforcing layout-invariant extraction. Finally, to guarantee clinical validity, the resulting corpus underwent expert-in-the-loop auditing by our clinical co-authors. This verification ensured that entity distributions, prose register, and structural formatting conventions strictly align with authentic medical records, filtering out any implausible instances.

[8] https://weasyprint.org

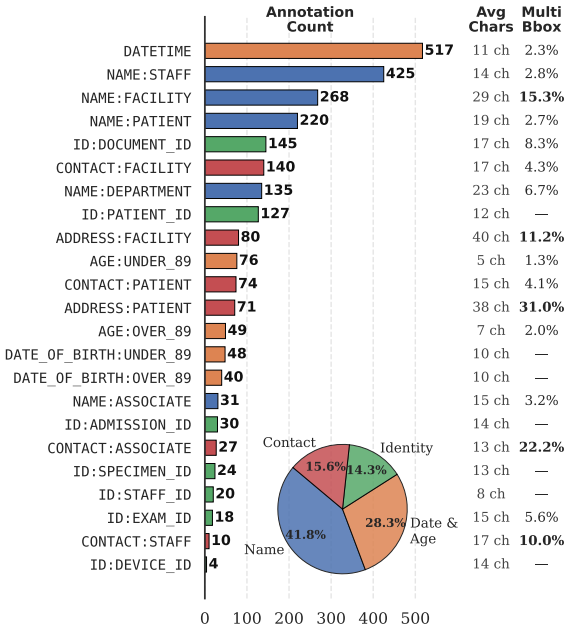


Figure 2: **Label distribution and structural complexity in PIB.** Right columns report average character length and layout fragmentation (bold indicates $\geq 10\%$).

3.3 Spatial Ground-Truth Annotation

To establish exact spatial grounding for each PHI instance, we developed a custom HTML interface (§A.7) for manual, page-by-page bounding-box labeling, compiling pixel-precise spatial coordinates into structured JSON ground truths to eliminate redundant padding around target tokens.

Importantly, when an entity span crosses line breaks, we enforce independent, disjoint bounding boxes rather than a single multi-row polygon. This prevents unannotated text within row margins from being caught in the crossfire of overlapping bounds.

Finally, prompts inform evaluated models that continuous spans may require multi-box pooling and that aggregated contacts require macro-level extraction, ensuring alignment with the layout design choices detailed in §3.1.

3.4 Inter-Annotator Agreement

To validate annotation quality, two annotators independently labelled each document following the guidelines in §A.7, with disagreements resolved by adjudication. We measure agreement on two axes: PHI label assignment and spatial grounding. For PHI label assignment, Cohen's $\kappa = 0.93$ indicates *almost perfect* agreement, confirming that the 23-category label taxonomy is unambiguous and that annotators share a consistent interpretation

of HIPAA entity boundaries. For spatial grounding, we compute the mean IoU between the two annotators' bounding boxes for matched entities, obtaining 0.87—well above the $\tau = 0.75$ threshold used in our strictest evaluation, indicating that pixel-level box placement is highly reproducible across annotators. Together, these figures establish that PIB rests on a reliable human signal, and that the evaluation thresholds we adopt are comfortably within the range of human agreement.

3.5 Labels Distribution

As visualized in Fig. 2, the dataset exhibits a pronounced long-tail distribution dominated by high-frequency administrative anchors such as `DATETIME` (n=517) and `NAME:STAFF` (n=425). Figure 2 also maps the schema's geometric footprint via average character length and layout fragmentation: spans range from dense tokens (`AGE:UNDER_89`, $\sim$5 chars) to lengthy narrative prose (`ADDRESS:FACILITY`, $\sim$40 chars), with 7.8% requiring multi-bbox spans due to text-wrapping, peaking for `ADDRESS:PATIENT` (31%) and `CONTACT:ASSOCIATE` (22.2%), structurally disadvantaging text-only sequence-labeling models.

Cross-domain topology reveals genre-specific shifts (Figure 3): administrative classes remain dense across genres, whereas peripheral entities are strictly localized (e.g., `ID:SPECIMEN_ID` is confined to lab workflows; 2.9% domain share, n=22). Further, PIB retains highly sparse tail categories (e.g., `ID:DEVICE_ID`, n=4; `CONTACT:STAFF`, n=10) to prevent majority-class shortcuts and stress-test model robustness against extreme class imbalance under localized geographic templates.

Page-level co-occurrence analysis reveals a deterministic layout topology: `DATETIME` tightly clusters with `NAME:STAFF` (137 pages) and `NAME:PATIENT` (135 pages). This administrative triad, coupled with `ID:DOCUMENT_ID` ($n \geq 129$ for all permutations), shows that parsing dense margins from narrative prose requires integrating global page-level context and visual spatial awareness.

3.6 Dataset Degradation

Evaluating on clean documents may yield optimistic robustness estimates, as model generalization can degrade when evaluated under conditions that differ from those observed during development (Moro et al., 2022); vision-language models suffer significant accuracy drops under real-world clinical artifacts (e.g., photocopying, faxing, scanning) (Cheng et al., 2025). To address this, each source page is replicated across three difficulty splits via a multi-stage augmentation pipeline built on Augraphy[9] that models artifacts at their physical source while preserving bounding boxes and PHI labels: **(1) Base:** unaugmented, clean 144 DPI (typical clinical document resolution) rasterizations serving as an ideal upper-bound reference. **(2) Medium:** activates nine conservative augmentations (e.g., ink bleed, lighting gradients, pen/highlighter annotations, scanner noise, subtle noise) to simulate hospital archives while maintaining legibility. **(3) Hard:** escalates medium effects and introduces three severe degradations—digital fax transmission distortions, physical fluid stains, localized lighting shadows (each sampled at 50% probability)—to probe model boundaries under extreme operational noise (Groleau et al., 2023).

All pages underwent manual review by our clinical co-authors; instances where severe artifacts rendered PHI text completely illegible were filtered out and regenerated, ensuring the benchmark evaluates extraction robustness under heavy noise rather than physically impossible inputs. This deterministic triplication expands the evaluation testbed to 414 total page instances. Samples of the three split are shown in §A.5 while detailed parameter specifications are provided in §A.4.

3.7 Dataset Structure

The final PIB benchmark is delivered as a structured dataset where each instance represents a single page mapped to a standardized schema. To maximize usability, the data format natively integrates both visual matrices and spatial annotations. Each page object contains the following root keys: `image` (the rendered page image matrix), `page` (the 1-based index within the source PDF), `total_pages` (the overall document length), `doc_type` (the domain identifier, e.g., `lab_reports`), `source_pdf` (the original source filename), and `annotations` (the array of PHI entities present on the page, which remains empty for PHI-free instances).

Each entity within the `annotations` array is formalized to support multi-box split spans through a structured dictionary containing: (i) `id`, a unique instance identifier; (ii) `label`, the hierarchical taxonomic string (`CATEGORY:SUBCATEGORY`); (iii) `text`, the literal UTF-8 ground-truth string;

[9] https://github.com/sparkfish/augraphy

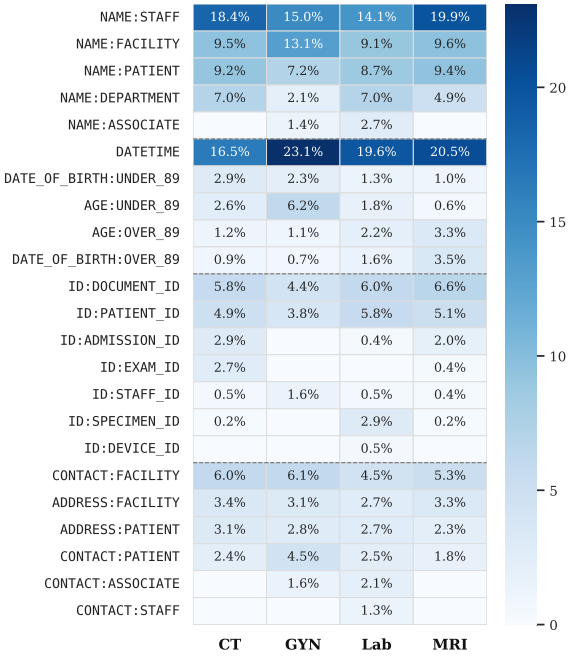


Figure 3: **Label distribution across document types.** Each cell reports the proportion of a category within its domain. dashed horizontal lines separate the macro-categories (Name, Date & Age, Identity, Contact).

(iv) `page_width/page_height` in PDF points; (v) `created`, an ISO 8601 generation timestamp; and (vi) `bboxes`, an array of spatial coordinates.

4 Experimental Setup

4.1 Task Formulation

Given a single page image, the model must return a JSON array in which each element specifies a PHI label from the 23-type controlled vocabulary (§3.1), the verbatim text string as printed in the image, and one or more bounding boxes $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ expressed as integers in $[0, 1000]$ normalized to page height and width, following the de facto standard for coordinate generation in MLLMs (Bai et al., 2023). Multiple boxes are permitted when an entity spans non-contiguous regions. An empty array signals a PHI-free page. No OCR stage or layout parser mediates the image-to-output path: a single forward pass must simultaneously read, localize, and classify every PHI token.

4.2 Benchmarked Models

Multimodal Models Data-governance constraints that prohibit routing clinical records through external APIs restrict our evaluation to models that can be deployed locally, with one closed-weight model included as a reference baseline. We selected recent state-of-the-art multimodal models available at the time of submission, covering both general-purpose and medically specialized systems, as well as dense and mixture-of-experts architectures. Our evaluation includes **Qwen3.5-9B**, **Qwen3.5-27B**, **Qwen3.5-122B-A10B**,[10] and **Qwen3.6-35B-A3B**,[11] representing recent native multimodal Qwen models spanning a broad range of model scales. We further evaluate **Gemma-4-31B** and **Gemma-4-26B-A4B** (Team et al., 2026), recent open-weight models designed for multimodal reasoning with long-context image understanding. **InternVL3.5-14B** (Wang et al., 2025) provides a complementary multimodal design and is particularly strong on document understanding, OCR, and visual grounding. Finally, **MedGemma-27B** (Research and DeepMind, 2025) represents a domain-specialized model, allowing us to assess whether medical-domain training translates into better performance on clinical visual understanding. As a closed-weight reference, we further include **GPT-5.4-mini**,[12] a recent multimodal model with strong reasoning capabilities, to provide a reference point beyond locally deployable open-weight systems.

OCR-based Models We additionally evaluate a more traditional **OCR+NER pipeline**, representing the conventional text-based approach to PDF de-identification. We use Docling[13] to extract page text and layout, followed by five zero-shot GLiNER (Zaratiana et al., 2024) models to identify PHI spans in the extracted text. Since GLiNER models are open-domain and accept natural-language entity labels, we map the PIB taxonomy to natural-language labels (§B.6). To accommodate their limited context windows, NER is performed at the sentence level. Model checkpoints, licenses, hardware specifications, and sampling configuration are given in §B.1, §B.2, §B.5.

4.3 Inference Protocol

All models receive the same prompt, consisting of a *system message* that injects the 23-label vocabulary with per-label descriptions (Table 12) and multi-box annotation guidelines, constraining output to a JSON array with no natural-language preamble,

[10] https://qwen.ai/blog?id=qwen3.5
[11] https://qwen.ai/blog?id=qwen3.6-35b-a3b
[12] https://developers.openai.com/api/docs/models/gpt-5.4-mini
[13] https://github.com/docling-project/docling

and a multimodal *user message* delivering a base64-encoded JPEG of the page image. Prompt template and output parsing details appear in §B.3 and §B.4.

4.4 Evaluation Metrics

End-to-end multimodal de-identification rests on three capabilities: *spatial grounding*—placing a bounding box that precisely covers the entity's page region; *mention extraction*—transcribing the entity's surface form verbatim; and *PHI type classification*—assigning the correct HIPAA label. A model can fail on any one while succeeding on the others: it may correctly localize and label an entity whose surface form it misread, transcribe a span verbatim while placing the box on the wrong line, or assign the wrong subtype to an entity it otherwise handled correctly. We report complementary metrics that isolate and combine these capabilities.

Before defining them, three conventions apply uniformly, except where noted. *Candidate filtering:* Only prediction–ground-truth pairs that share the same PHI label are considered for matching; a prediction with the wrong label is never a candidate and counts directly as a false positive. A ground-truth entity for which no same-label prediction is matched counts as a false negative, regardless of whether wrong-label predictions happen to overlap with it. *Greedy matching:* Each pair is assigned a similarity score—character-level F1 between normalized strings for text-based metrics, bounding-box IoU for spatial metrics. Pairs are accepted in descending score order, with each entity matched at most once; unmatched predictions count as false positives and unmatched ground-truth entities as false negatives. *True negatives:* True negatives are undefined; there is no notion of correctly predicting the absence of an entity at a specific location.

Each metric is annotated as $[\min, \max, \uparrow/\downarrow]$, where the arrow indicates whether higher ($\uparrow$) or lower ($\downarrow$) values are better; icons denote which capabilities are evaluated: T mention extraction (bbox-blind), spatial grounding (text-blind), both (no blindness). The complete metric suite, including additional diagnostic metrics for ablations, and all formal definitions are provided in §B.7.

T **Exact Span R / P / F1** $[0, 1, \uparrow]$ A candidate is a true positive if its normalized text (lowercased, whitespace-stripped) exactly matches the ground-truth string. A near-correct transcription counts as a false positive and a false negative at once (Sang and Meulder, 2003). Recall, precision, and their harmonic mean are macro-averaged across all 23 PHI label types so that each type contributes equally regardless of frequency.

Spatial F1 $[0, 1, \uparrow]$ A candidate is a true positive if its bounding box overlaps the ground-truth region by $\mathrm{IoU} \geq \tau$ (Everingham et al., 2010). We evaluate at $\tau \in \{0.25, 0.50, 0.75\}$ and report the arithmetic mean across all three thresholds, analogously to COCO mAP (Lin et al., 2014). Models with loosely overlapping boxes score well at $\tau = 0.25$ but degrade sharply at $\tau = 0.75$; robust models hold steady throughout the spectrum.

Entity Mean IoU $[0, 1, \uparrow]$ A complement to Spatial F1 that avoids binary IoU thresholding. For each ground-truth entity, we independently record the highest IoU achieved by any same-label prediction, or 0 if none exists; the average is taken over all of them.

T **Pass Rate** $[0, 1, \uparrow]$ A candidate is a true positive only when its bounding box overlaps the ground-truth region by $\mathrm{IoU} > 0.5$, and its normalized text exactly matches the ground-truth string. The gap between Pass Rate and either of the two bbox-blind or text-blind metrics above directly quantifies the additional cost of jointly requiring all three capabilities.

5 Results

Table 2 reports text, spatial, and joint Pass Rate (PR) performance across three difficulty splits. A clear dichotomy emerges: mention extraction and spatial grounding are empirically distinct capabilities that current MLLMs master to radically different degrees. An output example is shown in § B.12.

Text Performance Gemma-4-31B leads on average text F1 (73.7). However, Qwen3.5-27B surpasses it on the Hard split (70.9 vs. 68.3)—dropping only 2.7 points from Base to Hard compared to 8.4 points for Gemma-4-31B. MoE architectures consistently underperform their dense counterparts: Qwen3.5-122B-A10B trails the smaller Qwen3.5-27B, and Gemma-4-26B-A4B scores ∼4 F1 points below Gemma-4-31B. The closed GPT-5.4-mini reference model does not exceed this range: at 72.1 avg text F1, it closely trails the best open-weight models, indicating that frontier proprietary systems do not necessarily surpass open-weight extraction capabilities in this context.

	Base						Medium						Hard						AVG		
	T			T		T	T			T		T	T			T		T	T	T	T
Model	F1	P	R	F1	IoU	PR	F1	P	R	F1	IoU	PR	F1	P	R	F1	IoU	PR	F1	F1	PR
Multimodal Models																					
Gemma-4-31B-it	76.7	**82.1**	74.5	**58.7**	**54.2**	**63.5**	**76.0**	**80.6**	**74.5**	**58.0**	**53.9**	**62.6**	68.3	71.3	66.6	**57.1**	**53.2**	**58.8**	**73.7**	**57.9**	**61.6**
Gemma-4-26B-A4B-it	69.2	74.0	67.4	55.4	51.5	59.6	72.2	78.6	69.5	54.9	51.4	58.3	65.5	69.0	63.5	54.1	50.3	54.7	69.0	54.8	57.6
Qwen3.5-27B	73.6	79.2	70.8	39.8	38.4	38.7	73.1	75.8	73.2	40.8	41.3	38.6	**70.9**	**72.8**	**70.3**	40.2	40.4	35.9	72.5	40.3	37.7
Qwen3.6-35B-A3B	64.6	74.6	58.2	15.1	15.4	9.9	67.0	79.1	60.0	20.0	20.1	14.3	61.2	71.0	54.6	17.3	17.8	11.5	64.3	17.5	11.9
Qwen3.5-122B-A10B	74.3	77.4	74.4	10.6	12.4	7.4	72.5	75.7	72.5	17.4	19.6	13.0	67.5	68.6	67.6	15.5	18.0	10.3	71.4	14.5	10.2
Qwen3.5-9B	68.5	70.4	68.0	1.2	2.1	0.5	67.3	68.2	68.7	2.2	3.7	1.0	57.6	59.6	57.5	2.8	4.1	1.1	64.5	2.1	0.9
InternVL3.5-14B	55.6	67.2	50.6	0.0	0.3	0.0	50.3	61.0	45.4	0.0	0.3	0.0	44.5	53.1	40.1	0.0	0.4	0.0	50.1	0.0	0.0
MedGemma-27B-it	2.4	3.5	2.9	0.0	0.3	0.0	2.1	2.9	2.4	0.0	0.3	0.0	2.4	7.6	2.4	0.0	0.3	0.0	2.3	0.0	0.0
GPT-5.4-mini (high) 🔒	**77.6**	81.6	**76.5**	6.4	9.0	4.1	72.7	77.4	71.1	7.5	9.7	4.0	66.0	68.5	64.6	7.7	9.4	3.6	72.1	7.2	3.9
OCR-based Models																					
gliner2-privacy-filter-PII-multi	30.9	32.6	36.9	–	–	–	29.3	30.5	35.1	–	–	–	25.6	27.6	29.6	–	–	–	28.6	–	–
gliner-biomed-bi-large-v1.0	23.0	22.1	27.9	–	–	–	23.0	23.4	27.8	–	–	–	18.4	18.3	21.5	–	–	–	21.5	–	–
gliner_multi_pii-v1	20.4	30.7	21.3	–	–	–	20.8	29.5	21.8	–	–	–	16.1	25.4	16.9	–	–	–	19.1	–	–
gliner-PII	22.9	31.9	23.0	–	–	–	22.7	31.6	23.9	–	–	–	16.0	23.2	16.9	–	–	–	20.5	–	–
gliner-pii-large-v1.0	32.6	58.9	29.9	–	–	–	32.3	49.2	29.6	–	–	–	26.8	47.6	24.3	–	–	–	30.6	–	–

Table 2: **Overall performance on PIB.** Results are reported across three difficulty splits (Base / Medium / Hard) and their average (AVG). *Text* metrics: span-level macro F1, Precision, and Recall. *Spatial* metrics: Spatial F1 and Entity Mean IoU. *Joint* metric: Pass Rate. Open-weight models are sorted by decreasing average Pass Rate; GPT-5.4-mini, the only closed-source model evaluated, is reported last regardless of ranking. Best scores are reported in **bold**, while second-best scores are underlined, computed across all models.

Despite perfect adherence to the output format, MedGemma-27B collapses to 2.3 avg F1, suggesting the model fails to generalise to PHI extraction, likely due to overfitting on its training alignment rather than acquiring transferable extraction capabilities (more details are provided in § B.13). This finding is consistent with prior medical QA evidence that domain-specialized models are not necessarily more robust under controlled evaluation shifts and may exhibit even less safety-aligned abstention under uncertainty (Cocchieri et al., 2026c,d).

Spatial Performance Spatial grounding exposes an even steeper capability gap. Only the Gemma-4 dense models and Qwen3.5-27B produce meaningful boxes: Gemma-4-31B leads at 57.9 avg Spatial F1, Gemma-4-26B-A4B follows at 54.8, and Qwen3.5-27B reaches a moderate 40.3. All other models essentially fail: InternVL3.5-14B and MedGemma-27B both score 0.0 Spatial F1 despite being trained for spatial grounding, confirming that neither natural-image nor medical-image localisation generalises to dense document-text bounding. GPT-5.4-mini exhibits a similar drop-off: despite comparable text scores and high reasoning effort, it achieves just 7.2 avg Spatial F1. Critically, the Qwen MoE variants dissociate: Qwen3.5-122B-A10B achieves a strong text F1 of 71.4 yet a spatial F1 of only 14.5, showing that sparse activation preserves extraction ability while losing coordinate precision. Finally, spatial scores are more stable across difficulty splits than text scores: Gemma-4-31B drops only 1.6 Spatial F1 points from Base to Hard, versus 8.4 on text F1.

Pass Rate PR is the most demanding metric, requiring correct bounding box and verbatim text simultaneously. Gemma-4-31B leads at 61.6 avg PR, 12 points below its own text F1—a gap that directly quantifies the cost of adding the spatial requirement. For Qwen3.5-27B (PR 37.7), PR nearly matches its spatial F1 (40.3), indicating that spatial grounding is the binding constraint: when the box is correct, the transcription usually is too. The Qwen MoE models drop to a Pass Rate of 10–12 despite competitive text scores, confirming that their spatial failure is the decisive bottleneck under the joint metric. Models that fail spatially score zero PR by construction, making this metric a clean separator between models that have acquired end-to-end localisation and those that have not. Similarly, GPT-5.4-mini reaches only 3.9 avg PR (∼6% of Gemma-4-31B's score), suggesting the gap reflects intrinsic task difficulty rather than open-weight constraints.

Per-Domain Performance Model rankings are broadly stable across domains. **Gynaecology** consistently yields the highest text F1 (Gemma-4-31B-

	MRI						CT						Gynaecology						Laboratory					
	T			🖼		T🖼	T			🖼		T🖼	T			🖼		T🖼	T			🖼		T🖼
Model	F1	P	R	F1	IoU	PR	F1	P	R	F1	IoU	PR	F1	P	R	F1	IoU	PR	F1	P	R	F1	IoU	PR
Multimodal Models																								
Gemma-4-31B-it	72.3	**75.5**	70.5	**64.0**	**57.5**	**70.2**	**74.4**	76.4	**73.2**	**62.6**	**57.8**	**68.0**	**81.1**	**84.1**	79.3	**56.1**	**52.1**	**58.9**	**69.5**	**73.9**	**68.8**	**52.0**	**49.6**	**53.6**
Gemma-4-26B-A4B-it	67.6	70.1	66.0	61.1	56.3	66.7	72.7	**79.6**	69.4	61.3	56.3	66.3	75.9	80.3	73.5	53.3	49.1	54.9	63.7	67.6	62.9	47.1	45.3	47.3
Qwen3.5-27B	71.7	72.3	72.8	44.6	43.8	41.9	70.3	70.8	70.5	46.5	45.9	44.8	79.9	80.2	**80.0**	34.6	35.6	29.9	67.4	68.9	68.0	39.7	40.5	36.9
Qwen3.6-35B-A3B	61.5	72.1	55.0	15.6	15.5	9.9	66.7	73.6	61.8	23.0	23.5	18.0	72.3	80.6	66.4	14.5	15.5	9.0	57.9	71.7	50.7	17.1	16.9	11.2
Qwen3.5-122B-A10B	**75.0**	75.2	**76.9**	12.7	14.9	8.8	71.7	72.6	71.7	16.5	18.6	12.1	76.0	77.4	75.6	13.9	15.9	10.2	65.1	66.5	65.9	14.7	17.1	9.8
Qwen3.5-9B	65.6	65.8	67.2	2.1	3.3	0.9	65.4	68.3	64.1	2.7	4.3	1.1	68.4	68.0	71.0	1.7	2.8	0.7	58.8	61.3	58.9	2.0	3.0	0.9
InternVL3.5-14B	49.9	59.6	45.2	0.0	0.4	0.0	49.3	59.7	44.2	0.0	0.3	0.0	51.0	60.5	46.6	0.0	0.3	0.0	46.6	57.6	41.5	0.0	0.3	0.0
MedGemma-27B-it	1.9	2.1	2.4	0.0	0.3	0.0	3.5	4.8	4.2	0.0	0.3	0.0	1.0	1.0	1.2	0.0	0.3	0.0	3.3	7.4	3.3	0.0	0.3	0.0
GPT-5.4-mini (high) 🔒	68.6	70.3	68.2	7.0	9.4	3.7	71.2	72.8	70.7	8.9	11.0	5.3	80.0	82.0	78.6	7.8	8.5	3.3	63.5	66.4	62.4	6.4	8.9	3.5
OCR-based Models																								
gliner2-privacy-filter-PII-multi	30.0	36.2	36.5	–	–	–	35.7	38.1	38.6	–	–	–	28.8	29.4	35.2	–	–	–	25.2	28.5	28.4	–	–	–
gliner-biomed-bi-large-v1.0	23.0	23.5	24.4	–	–	–	28.3	32.4	28.8	–	–	–	21.2	21.3	25.5	–	–	–	19.0	22.5	19.3	–	–	–
gliner_multi_pii-v1	16.8	25.4	18.7	–	–	–	19.9	29.9	18.8	–	–	–	17.1	20.7	18.2	–	–	–	15.9	37.3	14.9	–	–	–
gliner-PII	18.2	29.5	20.4	–	–	–	23.1	30.5	22.2	–	–	–	17.4	24.1	20.9	–	–	–	14.2	26.8	14.5	–	–	–
gliner-pii-large-v1.0	33.7	59.7	30.4	–	–	–	31.4	43.7	30.0	–	–	–	25.5	38.4	26.0	–	–	–	26.4	48.0	21.8	–	–	–

Table 3: **Per-domain performance on PIB.** Results are reported across four document categories (MRI / CT / Gynecology / Laboratory), aggregated over all difficulty splits. *Text* metrics: span-level macro F1, Precision, and Recall. *Spatial* metrics: Spatial F1 and Entity Mean IoU. *Joint* metric: Pass Rate. Models are sorted by decreasing overall average Pass Rate (see Table 2). Best scores are reported in **bold**, while second-best scores are underlined.

it: 81.1; Qwen3.5-27B: 79.9), likely due to its more narrative and formulaic language. **Laboratory** is the hardest domain overall, with all models scoring lowest—or among the lowest—on text F1 and Pass Rate, suggesting that terse, value-dense reports challenge both extraction and grounding. Spatially, MRI and CT are more tractable: Gemma-4-31B-it reaches Spatial F1 of 64.0 and 62.6, respectively, versus 56.1 and 52.0 on Gynaecology and Laboratory, likely reflecting the more regular layouts of radiology reports. Notably, Qwen3.5-27B achieves its highest text F1 on Gynaecology (79.9) but its lowest Spatial F1 (34.6) on the same domain, showing that textual regularity does not guarantee spatial regularity. Fine-grained per-label results are reported in § B.9. For completeness, we additionally report three metrics: **Frequency-weighted Exact Span F1**, which weights entity occurrences equally, allowing frequent entity types to contribute proportionally to the score (§ B.8); **Transcription fidelity** (charF1 and exact match rate on text-matched pairs), which separates faithful PHI transcription from entity detection (§B.10); and **Spatial F1 across** $\tau \in 0.25, 0.50, 0.75$, which measures localization as the required bounding-box overlap becomes increasingly strict (§B.11).

OCR+NER Performance Since the OCR+NER pipeline does not involve bounding box extraction, we compare text metrics only. Its best model, gliner-pii-large-v1.0, achieves just 30.6 avg F1, a 2.4× drop from Gemma-4-31B, and trails every MLLM except MedGemma-27B (see Table 2). Unlike MLLMs, whose performance degrades gradually, GLiNER models tend to remain stable from Base to Medium but collapse on Hard. This sharp drop across all five variants points to OCR error propagation: Docling handles mild artifacts but fails under Hard degradations, producing corrupted or misaligned text that text-only models cannot recover. Consistent with Table 3, Laboratory is hardest for four of five GLiNER variants, likely due to dense tables and column misalignment, while CT is easiest for most due to its continuous prose.

6 Conclusion

We introduced PIB, the first benchmark for end-to-end, spatially grounded PHI de-identification in clinical PDFs, covering four document types, three visual degradation tiers, and a 23-category HIPAA-aligned taxonomy. Our evaluation reveals a sharp dichotomy: mention extraction is within reach of current MLLMs, while spatial grounding remains a decisive bottleneck not reliably addressed by domain-specific pre-training, grounding objectives, or architectural scale. No model approaches the reliability required for clinical deployment, and PIB is designed to track progress toward closing this gap. Future work should explore fine-tuning on dense text-level grounding data for robust spatial PHI localization, and extend PIB to multilingual records and more diverse clinical formats to better test generalization and real-world use.

Limitations

While our synthetic generation pipeline successfully bypasses the privacy hurdles associated with real-world medical data, it inherently remains an approximation of reality. The linguistic distribution of the generated prose depends heavily on the biases of the underlying LLM utilized during generation. Our current dataset configuration heavily biases towards English-speaking, limiting direct benchmarking for multi-lingual medical contexts. Furthemore, the rigorous manual bounding-box verification required for both clean and augmented splits introduces a significant human-effort bottleneck, reflecting the broader challenge of annotation efficiency in machine learning (Moro and Valgimigli, 2021), and limiting the rapid scaling of the dataset to tens of thousands of documents. Although PIB includes multi-page PDFs, evaluation is performed page by page and therefore does not capture cross-page context or document-level consistency. Future work should extend the benchmark to whole-document de-identification, potentially combining visual processing with semantic segmentation or hierarchical long-document modeling (Moro and Ragazzi, 2022). Finally, our evaluation focuses on effectiveness and does not account for the computational cost or environmental impact of the evaluated models. Future work should track inference-time energy consumption and carbon emissions alongside de-identification performance, following resource-aware evaluation frameworks such as Carburacy (Moro et al., 2023).

Risks and Ethical Considerations

PIB is built entirely from synthetic clinical documents: no real patient records, hospital records, or clinician identities were used at any stage of data collection or annotation. All medical scenarios and conditions were fabricated by the LLM-based generation pipeline and subsequently reviewed by practicing physicians solely to verify clinical plausibility, not to ground them in any real case.

LLM Memorization Risk The generative model used to produce the document text (Gemini-3-Flash) were pre-trained on large web corpora that may contain publicly accessible medical content such as anonymized case reports, physician directories, or hospital websites. It is therefore conceivable—though unlikely given the compositional, instruction-steered generation process—that the surface form of a name, institution, or location present in the training data could surface verbatim in a generated document. Nonetheless, users of PIB should be aware that absolute guarantees against memorized content are not achievable for LLM-generated text, and any downstream use that involves matching generated identifiers against external databases should treat the results accordingly.

Intended Use PIB is released exclusively for academic research into multimodal PHI de-identification. It should not be used as a substitute for real annotated clinical data in production de-identification systems, nor should models fine-tuned on PIB alone be deployed in clinical settings without additional validation on real, appropriately consented patient records. Since all documents are synthetic and contain no real patient data, PIB poses no re-identification risk.

Acknowledgements

Research partially supported by: AI-PACT (CUP B47H22004450008, B47H22004460001); National Plan PNC-I.1 DARE (PNC0000002, CUP B53C22006450001); PNRR Extended Partnership FAIR (PE00000013, Spoke 8); 2024 Scientific Research and High Technology Program, project "AI analysis for risk assessment of empty lymph nodes in endometrial cancer surgery", the Fondazione Cassa di Risparmio in Bologna; Chips JU TRISTAN (G.A. 101095947).

References

Kiana Aghakasiri, Noopur Zambare, JoAnn Thai, Carrie Ye, Mayur Mehta, J. Ross Mitchell, and Mohamed Abdalla. 2025. Not what the doctor ordered: Surveying LLM-based de-identification and quantifying clinical information loss. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 32187–32203. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Luca Beurer-Kellner, Marc Fischer, and Martin T. Vechev. 2024. Guiding LLMs the right way: Fast, non-invasive constrained generation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 3658–3673. PMLR / OpenReview.net.

Walter Cerroni, Gianluca Moro, Roberto Pasolini, and Marco Ramilli. 2015. Decentralized detection of network attacks through P2P data clustering of SNMP data. *Comput. Secur.*, 52:1–16.

Walter Cerroni, Gianluca Moro, Tommaso Pirini, and Marco Ramilli. 2013. Peer-to-peer data mining classifiers for decentralized detection of network attacks. In *Proceedings of the Twenty-Fourth Australasian Database Conference - Volume 137*, ADC '13, page 101–107, AUS. Australian Computer Society, Inc.

Alessio Cocchieri, Giacomo Frisoni, Marcos Martínez Galindo, Gianluca Moro, Giuseppe Tagliavini, and Francesco Candoli. 2025. OpenBioNER: Lightweight open-domain biomedical named entity recognition through entity type description. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 818–837. Association for Computational Linguistics.

Alessio Cocchieri, Giacomo Frisoni, Francesco Zangrillo, Luca Ragazzi, Marcos Martínez Galindo, Giuseppe Tagliavini, and Gianluca Moro. 2026a. OpenBioNER-v2: A suite of lightweight models for zero-shot medical named entity recognition via type descriptions. *Expert Syst. Appl.*, 318:131725.

Alessio Cocchieri, Luca Ragazzi, Gianluca Aguzzi, Giacomo Frisoni, Lorenzo Molfetta, Gianluca Moro, and Mirko Viroli. 2026b. Java academic benchmark: Exam-based evaluation of LLMs on object-oriented programming. *Journal of Systems and Software*, 242:113033.

Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, and Gianluca Moro. 2026c. LLMs (almost) never abstain under medical uncertainty. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 29573–29613. Association for Computational Linguistics.

Alessio Cocchieri, Luca Ragazzi, Giuseppe Tagliavini, and Gianluca Moro. 2026d. ReMedQA: Are we done with medical multiple-choice benchmarks? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 2706–2738. Association for Computational Linguistics.

Giacomo Domeniconi, Konstantinos Semertzidis, Vanessa López, Elizabeth M. Daly, Spyros Kotoulas, and Gianluca Moro. 2016. A novel method for unsupervised and supervised conversational message thread detection. In *DATA 2016 - Proceedings of 5th International Conference on Data Management Technologies and Applications, Lisbon, Portugal, 24-26 July, 2016*, pages 43–54. SciTePress.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. 2023. Augraphy: A data augmentation library for document images. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part III*, volume 14189 of *Lecture Notes in Computer Science*, pages 384–401. Springer.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Veysel Kocaman, Hasham Ul Haq, and David Talby. 2023. Beyond accuracy: Automated de-identification of large real-world clinical text datasets. *CoRR*, abs/2312.08495.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Gianluca Moro, Federico Di Luca, Davide Dardari, and Giacomo Frisoni. 2022. Human being detection from UWB NLOS signals: Accuracy and generality of advanced machine learning models. *Sensors*, 22(4):1656.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long documents in low-resource regimes. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11085–11093. AAAI Press.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14417–14425. AAAI Press.

Gianluca Moro and Lorenzo Valgimigli. 2021. Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature. *Sensors*, 21(19):6430.

Google Research and Google DeepMind. 2025. MedGemma technical report. *CoRR*, abs/2507.05201.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Praphul Singh, Charlotte Dzialo, Jangwon Kim, et al. 2025. RedactOR: An LLM-powered framework for automatic clinical data de-identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 510–530. Association for Computational Linguistics.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. *J. Biomed. Informatics*, 58:S20–S29.

Latanya Sweeney. 2002. k-Anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10(5):557–570.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, et al. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Qwen Team. 2025. Qwen3-VL technical report. *CoRR*, abs/2511.21631.

Jakov Vuković, Damir Ivanković, Claudia Habl, and Jelena Dimnjaković. 2022. Enablers and barriers to the secondary use of health data in Europe: General data protection regulation perspective. *Archives of Public Health*, 80.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, et al. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *CoRR*, abs/2307.09702.

Urchade Zaratiana, Ash Lewis, and George Hurn-Maloney. 2026. GLiNER2-PII: A multilingual model for personally identifiable information extraction. *Preprint*, arXiv:2605.09973.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: An efficient multi-task information extraction system with schema-driven interface. *CoRR*, abs/2507.18546.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5364–5376. Association for Computational Linguistics.

A Dataset Details

A.1 Document Generation Pipeline

Each document in the seed corpus is generated through a two-turn interaction with `gemini-3-flash-preview`. In the **first turn**, a *system message* establishes the rendering contract for the target document type, and a *user message* instantiates the concrete variant by injecting the parameters drawn from the combinatorial grid. In the **second turn**, a *user message* re-injects the PDF rendered by WeasyPrint into the model's multimodal context and requests targeted layout corrections without content alteration.

The following steps are executed independently for each parameter combination (see §A.2), hereafter referred to as a *variant*.

1. **Prompt assembly.** A type-specific Markdown system prompt encodes four categories of constraint: (i) the mandatory PHI entities that must appear in every document (Table 4); (ii) WeasyPrint-compatible HTML/CSS rules—all styling must be inline, page containers must be fixed-size `<div class="page">` elements at 210×297 mm, and no external resources are permitted; (iii) prose realism rules (flowing narrative, minor non-PHI typos, no bullet lists); and (iv) layout rules (near-monochrome palette, optional vertical margin separator, mandatory page footer). The user message serializes the variant's parameter tuple into a structured natural-language string that specifies country, patient sex and age range, document type–specific clinical parameters (exam type, lab panels, or clinical scenario), and the target visual style.

2. **HTML generation.** The assembled prompt pair is submitted to `gemini-3-flash-preview` via the Google Vertex AI Generative AI API within a stateful multi-turn chat session. The model returns a complete, self-contained HTML file. Any Markdown fences enclosing the response are stripped before further processing.

3. **Structural validation.** Two checks gate acceptance of the response: (i) the document must open with `<!DOCTYPE html>`, and (ii) the number of `<div class="page">` elements must exactly equal the expected page count for the seed type. Either failure triggers an immediate retry from step 2.

4. **PDF rendering and validation.** The validated HTML is rendered to a PDF via WeasyPrint with a global CSS override that enforces A4 dimensions and zero body margins. The resulting PDF page count is verified against the expected value; a mismatch triggers a retry.

5. **Visual self-correction (second turn).** The rendered PDF is re-injected into the same multimodal chat session. The model receives a fixed checklist of four admissible layout defects: (a) text clipped by or overlapping the page footer; (b) page content overflowing the page boundary; (c) excessive blank space between sections on multi-page documents; and (d) coloured sidebars not extending to the full page height. If any defect is identified, the model returns the complete corrected HTML; otherwise it responds with exactly the token `OK`. Every proposed correction is re-validated structurally and re-rendered before atomically replacing the accepted artefact. This step is configurable (`SELF_CORRECT_MAX` rounds) and is executed only after a successful initial render.

6. **Retry policy.** Steps 2–4 are retried up to three times with a 10-second inter-attempt backoff. A variant that exhausts all retries is recorded as *failed* in the manifest and does not produce output files.

7. **Artefact output.** Each successful variant produces three artefacts: an `.html` source file, a `.pdf` rendering, and a JSON record in a cumulative `manifest.json` that logs the full parameter tuple, observed page counts, self-correction rounds applied, and generation status.

A.2 Document Generation Parameter Grid

To ensure that the generated corpus covers a broad and controlled range of clinical, demographic, and visual conditions, each seed type is associated with a *parameter grid*: a set of named axes, each enumerating a discrete collection of values. Every variant in the corpus corresponds to one combination of values drawn from this grid, and the full corpus is obtained by sampling from—or exhaustively enumerating—the resulting combinatorial space. This structure serves two complementary roles. First, it provides a *systematic coverage guarantee*: by enumerating plausible values along each independent dimension, the grid ensures that no

PHI Category	Instantiation detail
Patient full name	Varied ethnicity; realistic English-language names
Date of birth	Full date or numeric (locale-appropriate format)
Age	Consistently derived from date of birth
Sex / Gender	Male or Female
National ID / SSN	SSN (US), NHS number (UK), or equivalent
Medical record number	Hospital-specific alphanumeric
Home address	Full address in locale-appropriate format
Phone number	Locale-appropriate mobile or landline
Email address	Next-of-kin contact only; omitted in ≈30% of documents
Requesting / referring physician	Full name, specialty, and title
Physician contact	Phone or email
Hospital / clinic name	Fictional but plausible
Hospital address	Full street address
Collection or exam date	Specific date and time
Document or sign-off date	Date and time
Responsible clinician	Full name and credential (MD, PhD, etc.)
Facility system ID	LIS/LIMS or RIS reference number
Next-of-kin / emergency contact	Full name, relationship, and phone

Table 4: **PHI categories mandated by the system prompt for every generated document.** All four seed types share these categories; individual types add further type-specific identifiers (e.g. specimen barcode for `lab`; exam code and radiologic technologist for `ct` and `mri`).

single demographic profile, institutional context, or rendering style dominates the corpus. Second, the combinatorial structure makes the benchmark extensible: new values can be appended to any axis independently, and the generation pipeline skips already-completed combinations via a manifest-based resume mechanism, so that an expanded run never re-generates existing artifacts.

Value Curation via Model-Assisted Enumeration and Human Validation The values populating each axis were not hand-crafted by the authors alone. For each axis, `gemini-3-flash-preview` was first prompted to generate an initial pool of diverse, clinically plausible candidates—covering representative exam types, common clinical indications, realistic laboratory panel compositions, and varied document layout conventions. The resulting candidates were then reviewed and curated by clinical co-authors to ensure medical realism, lexical diversity among candidate entries, and coverage of clinically meaningful edge cases (e.g., rare but realistic exam types, age-stratified pathology patterns, and country-specific institutional formatting conventions). The final value sets thus represent a human-validated, model-assisted enumeration of the realistic generation space for each seed type, a construction philosophy shared by other recent benchmarks pairing automated generation with expert human curation to ensure domain fidelity (Cocchieri et al., 2026b).

Axis

`country`. Drives locale-dependent formatting: date notation, phone patterns, national ID schemes (SSN, NHS number, Medicare card number, SIN), postal address conventions, and institutional naming norms across four English-speaking healthcare systems (US, UK, Australia, Canada).

`sex` / `age_range`. Controls demographic realism and ensures age-appropriate PHI patterns: paediatric vs. geriatric reference ranges, reproductive-age vs. post-menopausal clinical scenarios, and age-stratified background pathology. Nine age strata are shared across all seed types: teens, 20s, 30s, 40s, 50s, 60s, 70s, 80s, and 90s+.

`exam_type` / `panels` / `scenario`. Determines the domain-specific clinical scenario, directly shaping which PHI entities appear and in which textual contexts (e.g., an oncologist referral in a treatment-monitoring panel vs. a surgical referral in a pre-operative workup).

`indication` / `clinical_context`. Provides the referral narrative, varying the prose register and

the specific medical terms that surround PHI tokens.

`style`. The *visual diversity* axis and the most consequential for evaluating layout-robust PHI detection. For each seed type, a pool of natural-language layout descriptors was assembled spanning four document families: (i) *raw terminal outputs* (monospaced or plain-text, no typographic markup, resembling HIS / PACS / RIS system exports or voice-recognition transcription dumps); (ii) *formatted institutional reports* (colour header bar, bold or underlined section labels, bordered patient information tables); (iii) *margin annotation forms* (narrow left-column section codes separated from main content by a continuous vertical rule, as common in European hospital paper forms); and (iv) *fax and transmission printouts* (FAX routing block above the header, monochrome palette, transmission metadata lines). Documents from different families differ radically in visual rendering while preserving identical PHI coverage requirements, providing a stringent test of detector robustness.

The full cross-product of all axes constitutes the complete generation space. Tables 5 to 8 enumerate the grid for each seed type.

A.3 Prompt Templates

Each system message is organized into seven sections: (1) role and task definition; (2) document type and expected structure; (3) required PHI categories; (4) content variation rules; (5) writing style and realism rules—including the typographic noise distribution, contextual PHI embedding conventions, and dictation-style phrasing targets; (6) strict technical rules mandating WeasyPrint PDF compatibility; and (7) output format specification (raw HTML only, starting with `<!DOCTYPE html>`, no Markdown fences). All four seeds share this section architecture; the seed-specific content primarily concerns the document structure (sections 2 and 4), the required PHI identifiers (section 3), and the domain clinical constraints (section 5). Figs. 4 to 7 present a condensed rendering of the system message and the user message template for each seed. The complete prompt texts are publicly available in the project repository.

Parameter	Values / representative examples
`pages`	1 — urgent or limited panel, concise single-visit workup 2 — standard outpatient workup with 2–3 panels 3 — comprehensive workup with 4–5 panels and interpretive comments
`country`	US UK Australia Canada
`sex`	Male Female
`age_range`	teens 20s 30s 40s 50s 60s 70s 80s+ 90s+
`clinical_context` (12 values)	routine annual health check; pre-operative assessment; chronic kidney disease monitoring; diabetic management follow-up; suspected autoimmune disease workup; oncology treatment monitoring; suspected thyroid dysfunction; acute illness workup (fever and malaise); cardiovascular risk assessment; medication toxicity monitoring; post-surgical follow-up; infertility and hormonal workup
`panels` (12 values)	CBC + Clinical Chemistry + Lipid Profile; CBC + Liver Function Tests + Coagulation; CBC + Thyroid Panel + Inflammatory Markers; CBC + HbA1c + Renal Function + Urinalysis; CBC + Lipid Profile + Tumour Markers + Coagulation; CBC + Renal Function + Electrolytes + Urinalysis; CBC + Inflammatory Markers + Coagulation + Blood Cultures; CBC + Autoimmune Panel (ANA, ANCA, RF, anti-dsDNA) + Complement; CBC + HbA1c + Lipid Profile + Renal Function + LFTs; CBC + Thyroid Panel + Iron Studies + B12 + Folate; CBC + Toxicology Screen + LFTs + Renal Function; CBC + Bone Profile (Ca, PO_4, ALP, PTH) + Vitamin D + Renal Function
`style` (5 of ≈30)	[raw] central laboratory LIS printout, monospaced font, no header bar, analytes as plain text rows with inline reference ranges; [formatted] NHS laboratory report, Times New Roman, navy header bar, bold section headers, bordered patient info table; [margin] European laboratory form, narrow left-margin column with section codes (CBC, CHEM, LIP, COAG) and continuous vertical separator rule; [fax] fax-ready monochrome, FAX TRANSMISSION block above charcoal header, thin horizontal rules between sections; [fax] scanned pathology archive, serif font, faint institutional header, OCR-like spacing irregularities, archival photocopy aesthetic

Table 5: **Parameter grid for the clinical laboratory report seed (`lab`).** The `style` column shows 5 representative descriptors from a pool of ≈30 defined across four layout families (raw, formatted, margin, fax).

Parameter	Values / representative examples
`pages`	1 — single-region routine exam, concise findings 2 — multi-region or complex exam with detailed findings 3 — comprehensive staging or multi-system exam with extensive findings and prior comparison
`country`	US UK Australia Canada
`sex`	Male Female
`age_range`	teens 20s 30s 40s 50s 60s 70s 80s+ 90s+
`exam_type` (12 values)	CT Abdomen/Pelvis with and without IV contrast; CT Chest with contrast (PE protocol); CT Head without contrast; CT Urography with contrast; CT Neck with contrast; CT Chest/Abdomen/Pelvis with contrast (oncology staging); CT Pulmonary Angiography (CTPA); CT Abdomen without contrast (renal colic protocol); CT Cervical Spine without contrast; CT Thorax without contrast (HRCT); CT Coronary Angiography (CTCA); CT Aortography with contrast
`indication` (12 values)	acute abdominal pain, fever, and elevated inflammatory markers; sudden dyspnoea and pleuritic chest pain (suspected PE); severe headache, nausea, and visual disturbances; haematuria and flank pain (rule out urolithiasis or renal mass); neck mass and hoarseness in a smoker; known malignancy staging and treatment response assessment; acute chest pain with elevated D-dimer (rule out aortic dissection); blunt abdominal trauma after road traffic accident; chronic cough and weight loss (suspected primary lung malignancy); pre-operative workup for planned colorectal surgery; pulmonary nodule follow-up (Fleischner protocol); acute neurological deficit with sudden onset (rule out ischaemic stroke)
`style` (5 of ≈21)	[raw] plain institutional dictation output, no typographic treatment, patient info as colon-separated lines; [raw] voice-recognition transcription dump, monospaced font, no section labels, sections separated by blank lines only; [formatted] US academic medical centre EMR printout, compact patient banner, dark teal header bar, bold section headers; [margin] European radiology form, left-margin section tags (IND, TECH, FIND, IMP) aligned to content via continuous vertical separator rule; [fax] fax-ready monochrome CT report, FAX TRANSMISSION block above hospital header, plain bold ALL-CAPS section labels

Table 6: **Parameter grid for the CT radiology report seed (`ct`).** The `style` column shows 5 representative descriptors from a pool of ≈21 defined across four layout families (raw, formatted, margin, fax).

Parameter	Values / representative examples
`pages`	1 — single-region routine exam, concise technique and findings 2 — complex multi-sequence exam with detailed technique and findings 3 — comprehensive multi-sequence exam with prior comparison and management plan
`country`	US UK Australia Canada
`sex`	Male Female
`age_range`	teens 20s 30s 40s 50s 60s 70s 80s+ 90s+
`exam_type` (12 values)	MRI Pelvis without contrast (endometriosis protocol); MRI Brain without and with contrast; MRI Lumbar Spine without contrast; MRI Knee without contrast; mpMRI Prostate without contrast; MRI Cervical Spine without contrast; MRI Shoulder without contrast; MRI Abdomen with contrast (HCC surveillance protocol); MRI Breast with contrast (staging protocol); MRI Brain without contrast (epilepsy protocol); MRI Cardiac without and with contrast (function and viability); MRI Whole Spine without contrast (cervical + thoracic + lumbar)
`indication` (12 values)	chronic pelvic pain, dysmenorrhea, and deep dyspareunia (suspected recurrent endometriosis); progressive headache, memory impairment, and personality change; lower back pain radiating to left leg (suspected L4–L5 disc herniation); acute knee pain after sports injury (suspected meniscal or ligamentous tear); elevated PSA (8.2 ng/mL) on routine screening, no prior biopsy; neck pain with bilateral upper limb radiculopathy (suspected cervical myelopathy); shoulder pain and restricted range of motion after fall (suspected rotator cuff tear); known HCC, follow-up after locoregional treatment; BRCA1 mutation carrier, annual high-risk breast screening; first unprovoked generalised tonic-clonic seizure, new diagnosis workup; palpitations and dyspnoea, reduced ejection fraction on echocardiogram (cardiomyopathy); widespread neck and back pain with progressive lower limb weakness
`style` (5 of $\approx$31)	[raw] plain institutional MRI printout, serif font, dark grey header bar, all section labels as plain unformatted inline text, verbatim dictation feel; [raw] university research scanner text dump, fixed-width font, no header bar, patient and study metadata as plain comma-separated values; [formatted] NHS radiology letter, Times New Roman, navy header bar, bold underlined section headers, patient info in two-column bordered table; [margin] European radiology form, left-margin section codes (IND, TECH, FIND, IMP) aligned to content rows via continuous vertical separator rule; [fax] fax-ready monochrome MRI report, FAX TRANSMISSION block above hospital header, bold ALL-CAPS section labels, thin horizontal rules between sections

Table 7: **Parameter grid for the MRI radiology report seed (`mri`).** The `style` column shows 5 representative descriptors from a pool of $\approx$31 defined across four layout families (raw, formatted, margin, fax).

Parameter	Values / representative examples
pages	1 — brief follow-up visit or specialist results letter 2 — standard specialist consultation 3 — comprehensive first consultation
country	US UK Australia Canada
age_range	teens 20s 30s 40s 50s 60s 70s 80s+ 90s+
scenario(10 values)	uterine fibromatosis with menometrorrhagia, failed medical therapy; endometriosis stage III with dysmenorrhea, deep dyspareunia, and infertility; right ovarian endometrioma (4 cm), incidentally found on ultrasound; PCOS with oligomenorrhea, hyperandrogenism, and metabolic features; postmenopausal uterine bleeding, endometrial thickening 9 mm on transvaginal USS; primary infertility workup, suspected tubal factor, IVF counselling; anterior and posterior pelvic organ prolapse, grade II cystocele; adenomyosis with heavy menstrual bleeding, levonorgestrel IUS failure; high-grade cervical dyskaryosis (HSIL) follow-up after LLETZ procedure; premature ovarian insufficiency, secondary amenorrhoea aged 33, FSH 68 IU/L
style(5 of ≈23)	[raw] plain institutional printout, single-column, serif font, no typographic emphasis anywhere, patient info as raw colon-separated lines; [raw] colposcopy clinic system dump, monospaced font, no header bar, section labels as plain words in square brackets (e.g. [EXAMINATION]); [margin] European left-margin annotation column, section codes (IND, HX, EX, LAB, CONC, RX) aligned to content via continuous vertical separator rule, navy header; [sidebar] full-height left sidebar, navy background, department info panel in sidebar, sans-serif body, no bold section headers; [sidebar] full-height right sidebar, dark teal background, department info panel, sans-serif body, centred clinic header

Table 8: **Parameter grid for the gynaecological visit report seed (gynecology).** The style column shows 5 representative descriptors from a pool of ≈23 defined across four layout families (raw, formatted, margin, sidebar).

Prompt template — `lab` seed (clinical laboratory report)

System message:

Role and task. You are an expert medical document designer and HTML/CSS developer. Generate realistic synthetic clinical laboratory report variants in HTML format for a PHI de-identification benchmark.

Document structure. Hospital/lab header → patient info block → lab panel results tables (analyte | value | units | ref. range | flag) → interpretive comments → sign-off.

Required PHI (all 18 categories must appear): patient name (varied ethnicity), DOB, age, sex, national ID/SSN, MRN, home address, phone, next-of-kin email (*omitted in ≈30%*), requesting physician + specialty + contact, hospital/lab name + address, collection date/time, specimen barcode (alphanumeric), report validation date, validating pathologist + credential, LIS/LIMS system ID, next-of-kin/emergency contact.

Key content constraints.

- ≥30% of analytes flagged HIGH or LOW with realistic clinical deviations.
- Fax transmission block prepended in ≈20% of documents.
- Optional vertical left-margin annotation column with section codes in ≈10% of variants.
- Pages 2+: no repeated header or patient block; every page ≥80% full.
- Contextual PHI embedded in narrative: past surgeries (hospital + year), follow-up dates, specialist names.

Writing style and realism.

- Impression: dense paragraph of 2–5 sentences; **never** a bullet list.
- ≈60% of documents: 1–3 typographic errors in non-PHI, non-clinical narrative text only.
- PHI fields, medical terms affecting laterality/dosage, and clinical meaning: zero errors, always.

Mandatory WeasyPrint technical rules.

- Each page: `<div class="page">`, `width: 210mm; height: 297mm; overflow: hidden; page-break-after: always`.
- All CSS inline; no external resources; no `@page` rules; no JavaScript.
- Footer on every page: patient name · Page X of Y · validation line · document code.

User message template:

```
Generate a clinical laboratory report HTML variant:

- Pages: [1-3] ([page context description])
- Country setting: [US / UK / Australia / Canada]
- Patient sex: [Male / Female]
- Patient age range: [e.g. teens, 40s, 70s, 90s+]
- Clinical context: [e.g. routine annual check, CKD monitoring, oncology treatment]
- Lab panels: [e.g. CBC + HbA1c + Renal Function + Urinalysis]
- Visual style: [style descriptor from grid]
```

Figure 4: **Condensed system prompt and user message template for the `lab` seed.** The system message encodes 18 mandatory PHI categories, panel-level result formatting rules, prose realism constraints, and strict WeasyPrint rendering rules. The user message instantiates the concrete variant by injecting the parameter tuple drawn from the combinatorial grid (Table 5).

Prompt template — `ct` seed (CT radiology report)

System message:

Role and task. You are an expert medical document designer and HTML/CSS developer. Generate realistic synthetic CT radiology report variants in HTML format for a PHI de-identification benchmark.

Document structure. Hospital/department header → patient info block → exam details (exam type, exam code, date, clinical indication) → technique *(optional)* → findings → impression → sign-off.

Required PHI: patient name (varied ethnicity), DOB, sex, home address, phone, patient ID, admission ID, exam code, exam date, referring department + physician, reporting radiologist, radiologic technologist, hospital name + address + contact.

Key content constraints.

- Findings: continuous organ-by-organ prose (≈200 words per page of findings); no bullet lists or per-organ sub-headings.
- Impression: dense paragraph (2–5 sentences); **must not** reference other specialists, management plans, or scheduled procedures.
- No "Plans", "Recommendations", or "Plan & Recommendations" section — these do not exist in real radiology reports.
- Findings continuation across pages: picks up mid-sentence; no "FINDINGS (CONTINUED)" or any label.
- Fax transmission block in ≈20% of documents; dictation-style opening of Findings in ≈30%.
- Pages 2+: no repeated header; every page ≥80% full.

Writing style and realism. Standard radiological language; at least one positive finding plus normal structures. ≈60% of documents: 1–3 typographic errors in non-PHI narrative text only. PHI fields and laterality/dosage terms: zero errors.

Mandatory WeasyPrint technical rules. Same as `lab` seed (fixed A4 `<div class="page">` containers, inline CSS, pinned footer with patient ID and validation line).

User message template:

```
Generate a CT radiology report HTML variant:

- Pages: [1-3] ([page context description])
- Country setting: [US / UK / Australia / Canada]
- Patient sex: [Male / Female]
- Patient age range: [e.g. 20s, 50s, 80s+, 90s+]
- CT exam type: [e.g. CT Abdomen/Pelvis with contrast, CT Chest PE protocol]
- Clinical indication: [e.g. acute abdominal pain + fever, suspected PE]
- Visual style: [style descriptor from grid]
```

Figure 5: **Condensed system prompt and user message template for the `ct` seed.** The impression is explicitly constrained to imaging findings only, with no management recommendations or specialist referrals, mirroring real radiological reporting conventions. The user message injects the parameter tuple from Table 6.

Prompt template — `mri` seed (MRI radiology report)

System message:

Role and task. You are an expert medical document designer and HTML/CSS developer. Generate realistic synthetic MRI radiology report variants in HTML format for a PHI de-identification benchmark.

Document structure. Hospital/department header → patient info block → exam details → **technique** (MRI sequences, planes, contrast agent or none) → findings → impression → sign-off.

Required PHI: same as `ct` seed, with radiologic technologist replaced by MRI technologist; exam code omitted; admission ID added.

Technique realism (sequence set must match the exam type):

- Endometriosis pelvis: axial/sagittal/coronal T2, T1, T1 fat-sat, DWI — no contrast.
- Brain with contrast: T1, T2, FLAIR, DWI, ADC, T1 post-Gd.
- Brain epilepsy: coronal T2, FLAIR, T1 3D MPRAGE, hippocampal volumetry — no contrast.
- Lumbar/cervical spine: sagittal T1 + T2 + STIR, axial T2 at key levels — no contrast.
- mpMRI prostate: T2 multiplanar, DWI/ADC, DCE.
- Liver/abdomen: T1 in/out-of-phase, T2 fat-sat, DWI, dynamic post-Gd (arterial/portal/delayed).
- Cardiac: cine SSFP, T2 STIR, LGE post-Gd, T1/T2 mapping.

Key content constraints.

- Findings: continuous prose with inline bold compartment labels ($\approx$220 words per page); no bullet lists.
- Impression: same restrictions as `ct` — imaging findings only; no management referrals.
- Header bars: compact ($\leq$10 mm height), monochrome only — no large coloured banners.
- Findings continuation across pages: picks up mid-sentence; no continuation label.
- Pages 2+: no repeated header; every page $\geq$80% full.

Writing style and realism. Same typographic noise distribution and dictation conventions as `ct` seed. MRI-specific abbreviations: T1WI, T2WI, STIR, DWI, ADC, DCE, Gd, bilat., approx., c/w, s/p.

Mandatory WeasyPrint technical rules. Same as `lab` seed.

User message template:

```
Generate an MRI radiology report HTML variant:

- Pages: [1-3] ([page context description])
- Country setting: [US / UK / Australia / Canada]
- Patient sex: [Male / Female]
- Patient age range: [e.g. teens, 30s, 60s, 90s+]
- MRI exam type: [e.g. MRI Pelvis without contrast, mpMRI Prostate]
- Clinical indication: [e.g. chronic pelvic pain / suspected endometriosis]
- Visual style: [style descriptor from grid]
```

Figure 6: **Condensed system prompt and user message template for the `mri` seed.** The technique section requires the model to select and enumerate the correct MRI pulse sequences for the requested exam type, grounding the generated document in protocol-level clinical realism. The user message injects the parameter tuple from Table 7.

Prompt template — gynecology seed (gynaecological visit report)

System message:

Role and task. You are an expert medical document designer and HTML/CSS developer. Generate realistic synthetic gynaecological outpatient specialist consultation and follow-up report variants in HTML format for a PHI de-identification benchmark.

Document structure (content depth scales with page count):
Multi-page variants: hospital/dept header → report title → patient info block → clinical indication → full medical history (menstrual, obstetric, medications, screening dates, BMI/height/weight, family Hx) → examination findings → conclusions/management plan → sign-off.
1-page variants: compact header (2 lines max), single-line patient block, brief background (1–2 sentences), key findings (one prose block), conclusions + plan (one short paragraph), sign-off. Full menstrual/obstetric history and screening lists are **omitted**.

Required PHI: patient name (varied ethnicity), DOB, age, patient ID, exam date, referring physician, home address, phone, **patient email**, BMI/height/weight, LMP (specific date), screening test dates + results (HPV, Pap smear, mammography), reporting physician + specialty, hospital name + address + contact; emergency contact (≈*30% of documents*).

Key content constraints.

- Examination findings: continuous prose; uterus dimensions (3-axis mm) + position + texture; pathological structures with measurements; each ovary; Pouch of Douglas; Doppler if relevant.
- Conclusions: dense paragraph including future appointment date and named follow-up physician.
- Patient referred to as "the patient", "she", "her" in all narrative text; name appears *only* in the patient info block and the page footer.
- Contextual PHI in narrative: past surgeries with hospital name + year; planned procedure dates; specialist names.
- Pages 2+: no repeated header; every page $\geq$80% full.

Writing style and realism. Dictation register; gynaecological abbreviations (LMP, G2P2, PCOS, HRT, HPV, USS, OCP). ≈60% of documents: 1–3 typographic errors in non-PHI narrative text only.

Mandatory WeasyPrint technical rules. Same as `lab` seed.

User message template:

```
Generate a gynaecological visit report HTML variant:

- Pages: [1-3] ([page context description])
- Country setting: [US / UK / Australia / Canada]
- Patient age range: [e.g. 20s, 40s, 60s, 90s+]
- Clinical scenario: [e.g. endometriosis stage III, PCOS, postmenopausal bleeding]
- Visual style: [style descriptor from grid]
```

Figure 7: **Condensed system prompt and user message template for the gynecology seed.** This seed introduces two distinctive constraints absent from the radiology seeds: (i) the patient's name is banned from narrative text and may appear only in the info block and footer, increasing the difficulty of PHI localisation; (ii) the 1-page variant has a reduced document structure that excludes full medical and screening histories. The user message injects the parameter tuple from Table 8.

A.4 Augmentation Pipeline

The three difficulty splits are produced by a stochastic, multi-stage augmentation pipeline built on Augraphy. This section documents the complete per-augmentation parameter specifications for the medium and hard splits; the base split is unaugmented and requires no further specification.

The pipeline enforces a fixed three-phase execution order that mirrors the physical sequence in which degradation accumulates prior to digitization. The *ink phase* (InkBleed, LowInk, InkMottling, LinesDegradation) models degradation originating at the toner or ink layer. The *paper phase* (ColorPaper) models substrate-level aging (yellowing, tinting). The *post phase* (LightingGradient, SubtleNoise, Stains, ShadowCast, Annotations, Faxify, ScannerNoise) models artifacts introduced during scanning, physical handling, and clinical annotation. Phase ordering is fixed; augmentation order within each phase follows the top-to-bottom sequence in Tables 9 and 10.

Table 9 specifies the medium split, in which nine augmentations are active at conservative intensity levels designed to degrade signal without impairing readability: the warmest ColorPaper tint produces only a mild amber shift, and the sampling range of number of annotations is lower. Notably, ScannerNoise is applied exclusively in the medium split at conservative intensity, as it tends to compound destructively with the severe augmentations active in the hard split, producing samples that exceed our legibility threshold. As a result, the medium split is not merely a milder version of the hard split, but follows a partially distinct augmentation trajectory. Most parameters are sampled independently and uniformly at random within the stated ranges for each page; augmentations marked † in the table are *profile-sampled*: one discrete archetype—a named bundle of correlated parameter values—is committed to at the start of the pipeline and held fixed for that page, ensuring internally coherent and valid degradation styles.

Table 10 specifies the hard split, which escalates all medium parameters (except ScannerNoise) and introduces three additional severe degradations. These new degradations are subject to a conditional activation rule indicated by ‡ in the table, which is a separate mechanism from profile sampling: when Faxify fires (probability 0.5), Stains and ShadowCast are suppressed for that page, preventing physically implausible combinations. The Faxify augmentation is the most disruptive single component: its binarization collapses the entire grayscale spectrum into a two-tone output, rendering all downstream ink and paper effects invisible, and making character boundary recovery dependent solely on the surviving high-contrast halftone pattern. The Annotations augmentation—active at 50% probability—places 35–60 synthetic pen or highlighter strokes per page in the hard split, directly occluding PHI regions and simulating the clinician markup that is ubiquitous in paper-based hospital workflows.

Following augmentation, every page undergoes manual review by our clinical co-authors against the original bounding-box annotations; instances in which any PHI token is rendered illegible are rejected and regenerated, ensuring that all splits remain solvable by a trained human reviewer and that the benchmark evaluates robustness under heavy noise rather than physically impossible inputs.

A.5 Benchmark Samples

To make the visual effect of the three difficulty splits immediately comparable, Fig. 8 renders the same source page under each degradation tier side-by-side, holding instance content constant so that any perceptual difference is attributable solely to the augmentation pipeline. The base split (Fig. 8a) provides the upper-bound legibility reference: 144 DPI rasterization with no augmentation, sharp glyph edges, uniform white background, and full color fidelity. The medium split (Fig. 8b) introduces moderate noise—warm amber paper tint (ColorPaper), a top-to-bottom luminance gradient (LightingGradient), and mild ink spread along high-contrast strokes (InkBleed)—while keeping every PHI token fully legible, faithfully simulating a good-quality hospital archive scan. The hard split (Fig. 8c) escalates all medium effects and compounds them with severe degradations: Faxify binarization collapses grayscale gradients into high-contrast halftone patterns, InkBleed and LinesDegradation fracture individual character strokes, and elevated LowInk banding erodes uniform fills, collectively replicating a document that has traversed fax and photocopy cycles before scanning. Bounding-box annotations and PHI labels are held fixed across all three renderings; the splits therefore probe model robustness in isolation from annotation quality.

Output A.1 reproduces five representative entities from page 1 of the same document shown in

Fig. 8 (the complete annotation comprises 23 entities across two pages). NAME:FACILITY spans two disjoint bounding boxes because the facility name wraps across a line break in the document layout.

Output A.1: Ground-truth annotation excerpt


```
{
  "sample_0033_lab_us_female.pdf": {
    "pages": { "1": [
      { "label": "NAME:FACILITY",
        "text":  "WESTSHORE REGIONAL MEDICAL
    CENTER",
        "bboxes": [
          [0.07237, 0.10399,
          0.08564, 0.55822],
          [0.09062, 0.10399,
          0.10390, 0.21197]] },
      { "label": "NAME:PATIENT",
        "text":  "WHITMORE, ELEANOR GRACE",
        "bboxes": [[0.17776, 0.09577,
        0.18772, 0.32700]] },
      { "label": "DATE_OF_BIRTH:UNDER_89",
        "text":  "07/12/1951",
        "bboxes": [[0.17859, 0.37042,
        0.18855, 0.45141]] },
      { "label": "AGE:UNDER_89",
        "text":  "72Y",
        "bboxes": [[0.17720, 0.46000,
        0.18772, 0.48897]] },
      { "label": "CONTACT:FACILITY",
        "text":  "(941) 555-0188",
        "bboxes": [[0.12132, 0.43146,
        0.13180, 0.54296]] }
      // ...
    ] } }
}
```


A.6 Annotation Interface

Producing pixel-precise bounding-box annotations for 414 page instances across three difficulty splits requires a tooling environment that satisfies three simultaneous constraints: native PDF rendering without format conversion, strictly local data processing to comply with the HIPAA privacy requirements, and native support for the multi-box span structure formalized in §3.7. Existing annotation platforms satisfy at most two of these constraints: tools that offer native PDF rendering (e.g., Encord, UBIAI) operate exclusively as cloud SaaS platforms, requiring data upload to external servers; fully local tools (e.g., Label Studio in offline mode) handle text-box annotation but lack native multi-page PDF rendering and sub-pixel spatial precision; and none natively support the disjoint multi-box encoding that line-wrapped PHI spans require. We therefore implement a self-contained, browser-native annotation interface delivered as a single HTML file loaded locally, requiring no network connectivity or server infrastructure. The complete annotation interface is publicly released as part of the PIB codebase, enabling the community to reproduce our annotations, extend the benchmark to new document types, or adapt the tool for other spatially grounded NLP tasks.

The tool ingests a folder of PDF documents via the File System Access API and renders each page at configurable resolution through PDF.js.[14] An overlay canvas exposes bounding-box drawing by click-and-drag; upon releasing the mouse, a modal form requires the annotator to select one of the 23 canonical PHI labels from a grouped drop-down and to transcribe the verbatim surface form visible within the box before the annotation can be committed, enforcing completeness at entry time. For entities whose surface forms cross line breaks, the interface supports multi-box mode: additional disjoint boxes are appended to the same annotation record, directly encoding the non-contiguous span structure formalized in § 3.7. A sub-pixel fine-tuning widget enables nudging each box edge by ± 1 or ± 5 canvas pixels, allowing annotators to achieve the pixel-level placement precision required for tight IoU scores at the $\tau = 0.75$ evaluation threshold. A fixed-box stamp mode, configurable per label, accelerates annotation of spatially regular fields such as dates and patient identifiers by centering a preset-size box on the click point, reducing per-entity annotation time for the most frequent PHI categories.

A.7 Annotation Guidelines

The guidelines below were distributed to all annotators before the labelling session and kept open as a reference throughout. Disagreements between the two independent annotators were resolved by adjudication against these rules (Fig. 12).

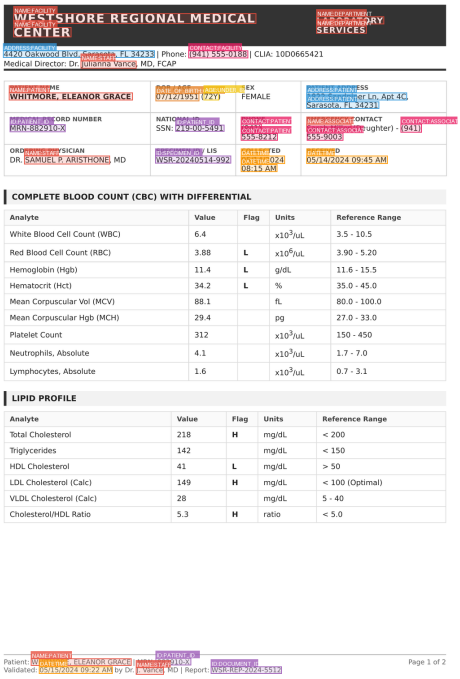
[14] https://github.com/mozilla/pdf.js

Phase	Augmentation	p	Parameter ranges
Ink	InkBleed	1.00	intensity $[0.70, 1.00]$; kernel $[3, 6]$; severity $[0.38, 0.85]$
	InkMottling	0.50	alpha $[0.12, 0.20]$; noise scale $[2, 6]$; Gaussian kernel $[3, 11]$
	LowInk	1.00	random bands $[3, 8]$; periodic bands $[3, 6]$, period $[3, 8]$
	LinesDegradation	1.00	gradient $[32, 255]$; split probability $[0.2, 0.6]$
Paper	ColorPaper†	0.70	2 profiles, weights $[0.5, 0.5]$; hue $[18, 50]$; saturation $[5, 20]$ (profile 1), $[15, 35]$ (profile 2)
Post	LightingGradient	1.00	Gaussian mode; transparency 0.90; min. brightness 50
	SubtleNoise†	1.00	2 profiles: r=25 (wt.= 0.5), r=35 (wt.= 0.5)
	ScannerNoise	0.70	bad-photocopy (type 1); noise value $[60, 100]$; sparsity $[0.3, 0.7]$; concentration $[0.2, 0.6]$; iter $[1, 3]$
	Annotations†	0.50	highlight markup; pen/highlighter ink; 5–20 strokes per page; 4 highlight colours

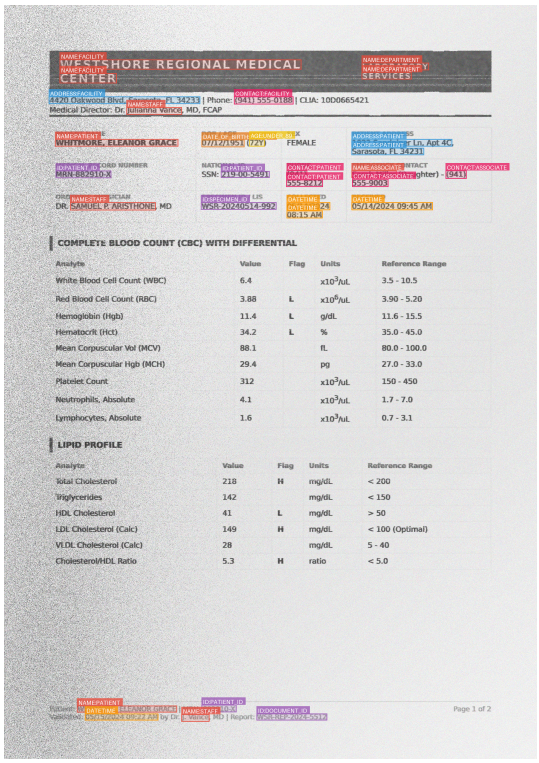
Table 9: **Augmentation parameters for the *medium* split.** p = per-page activation probability; all numeric ranges are sampled uniformly. Augmentations absent from this table are disabled ($p = 0$). † = profile-sampled: one archetype committed per page before execution.

Phase	Augmentation	p	Parameter ranges
Ink	InkBleed	1.00	intensity $[0.65, 0.90]$; kernel $[3, 8]$; severity $[0.40, 0.70]$
	InkMottling	0.50	alpha $[0.28, 0.50]$; noise scale $[2, 8]$; Gaussian kernel $[5, 21]$
	LowInk	1.00	random bands $[3, 8]$; periodic bands $[3, 6]$, period $[3, 8]$
	LinesDegradation	1.00	gradient $[32, 255]$; split probability $[0.2, 0.6]$
Paper	ColorPaper†	1.00	3 profiles, weights $[0.2, 0.4, 0.4]$; saturation up to $[40, 80]$
Post	LightingGradient	1.00	Gaussian mode; transparency 0.90; min. brightness 50
	SubtleNoise†	1.00	2 profiles: r=18 (wt.= 0.5), r=30 (wt.= 0.5)
	Annotations†	0.50	highlight markup; pen/highlighter ink; 35–60 strokes per page; 4 highlight colours
	Stains†‡	1.00	2 profiles: *fine_stains* (wt.= 0.4, normal blend, α=0.2); *light_stains* (wt.= 0.6, darken blend, α=1.0)
	ShadowCast‡	0.50	width $[0.05, 0.80]$; opacity $[0.05, 0.60]$; soft blur $[51, 401]$ px
	Faxify†‡	0.50	3 modes: mono-only (wt.= 0.4), halftone (wt.= 0.4), both (wt.= 0.2); threshold algorithms: Li, mean, Otsu, Sauvola

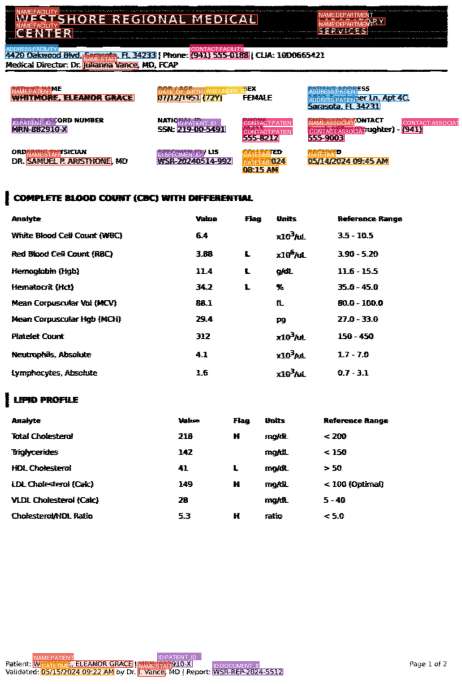
Table 10: **Augmentation parameters for the *hard* split.** † = profile-sampled: one archetype committed per page before execution. ‡ = mutually exclusive pair: Faxify suppresses Stains and ShadowCast.



(a) **Base split.** Clean 144 DPI rasterization.



(b) **Medium split.** ColorPaper tint, LightingGradient, and InkBleed.



(c) **Hard split.** Faxify binarization, InkBleed, Lines-Degradation, and LowInk banding.

Figure 8: **PIB difficulty splits on a representative page.** All three renderings correspond to the same source document (`sample_0033_lab_us_female.pdf`, page 1); bounding-box annotations and PHI labels are invariant across splits. The progression from base to hard mirrors the physical sequence of degradations that accumulate in real hospital document workflows. Text remains fully legible.

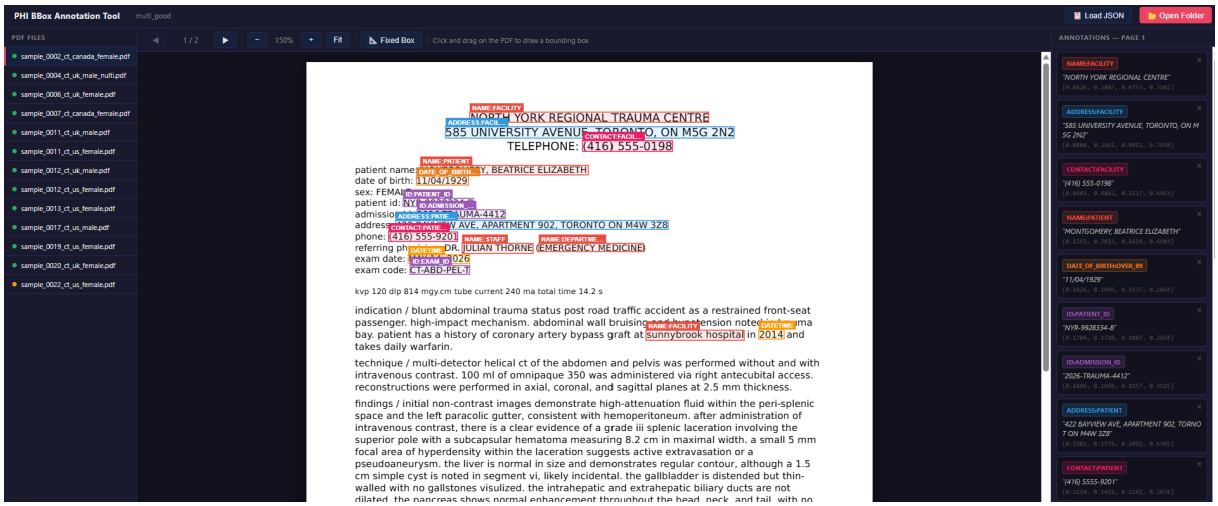


Figure 9: **Screenshot of the annotation tool interface.** The left panel shows the document file-list with per-document completion status. The central panel renders the PDF page, with drawn bounding boxes overlaid and colour-coded by PHI label. The right panel shows the structured annotation store for the current page, listing label, verbatim surface form, and normalized bounding-box coordinates for each committed entity.

B Experimental Details

B.1 Model Checkpoints and Licensing

Canonical HuggingFace checkpoint identifiers and license terms for all evaluated models are listed in Table 11. Eleven of the thirteen models are released under the Apache 2.0 license. The exceptions are MedGemma-27B, distributed under the Google Health AI Developer Foundations terms of use and gliner-PII, under NVIDIA Open Model License Agreement, that remain compatible with the on-premise deployment scenario PIB targets, though both introduce additional legal compliance overhead relative to the Apache 2.0 majority—a distinction that practitioner adopters should factor into planning.

B.2 Inference Hyperparameters

All models employ stochastic decoding and are served at provider defaults unless a specific configuration file overrides them. For models that perform chain-of-thought reasoning (Wei et al., 2022) prior to emitting the structured output—identifiable by the presence of `<think>` blocks in their responses—temperature is set to a strictly positive value following developer recommendations, as temperature zero induces degenerate repetition in extended reasoning traces. The maximum generation budget is set high enough to accommodate both the full thinking trace and the JSON output, the latter of which can itself exceed several hundred tokens on annotation-dense pages. All hyperparameters can be overridden by edit YAML configuration files distributed alongside the benchmark, ensuring that any future evaluation extending PIB to additional models can be conducted under identical conditions without modifying inference code. The closed-weight reference model, GPT-5.4-mini, is queried at *high* reasoning effort, trading additional inference-time compute for extraction accuracy. GLiNER-based models are evaluated at each checkpoint's default entity-extraction confidence threshold (0.5), without per-model or per-label tuning, to keep the OCR+NER baseline comparison uniform across variants.

B.3 Inference Prompt Templates

Fig. 10 reproduces verbatim the system and user turns submitted to all evaluated models. For MedGemma-27B and InternVL3.5-14B, we slightly adapted the system prompt following the recommendations provided by the respective model authors; the exact templates are available in our code repository. The `{labels_with_guidelines}` placeholder is substituted at runtime with the comma-separated enumeration of all 23 canonical label strings from the PIB taxonomy, each paired with its natural-language description; the complete substitution is listed in Table 12. Injecting per-label descriptions directly into the context—rather than relying on label strings alone—reduces label-confusion errors on rare entity types without any parameter update (Cocchieri et al., 2025, 2026a). Coordinates are requested as integers in

Model	HuggingFace Checkpoint	Licence
Qwen3.5-9B	Qwen/Qwen3.5-9B	Apache 2.0
Qwen3.5-27B	Qwen/Qwen3.5-27B	Apache 2.0
Qwen3.5-122B-A10B	Qwen/Qwen3.5-122B-A10B	Apache 2.0
Qwen3.6-35B-A3B	Qwen/Qwen3.6-35B-A3B	Apache 2.0
Gemma-4-31B	google/gemma-4-31B-it	Apache 2.0
Gemma-4-26B-A4B	google/gemma-4-26B-A4B-it	Apache 2.0
InternVL3.5-14B	OpenGVLab/InternVL3_5-14B	Apache 2.0
MedGemma-27B	google/medgemma-27b-it	Health AI Developer Foundations
gliner2-privacy-filter-PII-multi	fastino/gliner2-privacy-filter-PII-multi	Apache 2.0
gliner-biomed-bi-large-v1.0	Ihor/gliner-biomed-bi-large-v1.0	Apache 2.0
gliner_multi_pii-v1	urchade/gliner_multi_pii-v1	Apache 2.0
gliner-PII	nvidia/gliner-PII	nvidia-open-model-license
gliner-pii-large-v1.0	knowledgator/gliner-pii-large-v1.0	Apache 2.0

Table 11: **HuggingFace checkpoints and licences for all benchmarked models.** URLs resolve to the canonical model card on https://huggingface.co.

$[0, 1000]$ rather than as floating-point values in $[0, 1]$: this matches the convention adopted by most modern MLLMs, which are trained to predict bounding boxes in this normalized integer format (Team, 2025). Coordinates are rescaled back to $[0, 1]$ during output normalization (§ B.4). A maximum quantization error of 1 unit in $[0, 1000]$ space corresponds to 0.001 in normalized coordinates, or at most $0.001 \times \max(\text{height}, \text{width}) = 0.001 \times 1{,}684 \approx 1.7$ pixels on an A4 page rasterized at 144 DPI.

B.4 Output Parsing and Normalization

Before any metric is computed, model replies pass through a deterministic four-step normalization pipeline that ensures a consistent evaluation surface regardless of inference backend or model family.

Step 1 — Chain-of-Thought Stripping Models that perform explicit reasoning prior to structured output (e.g., Qwen3.5/3.6 variants) enclose their thinking traces inside `<think>...</think>` delimiters before emitting the final response. These blocks are stripped in full before JSON parsing. Some inference backends—notably vLLM when its chat template prepends `<think>` to the prompt to enforce reasoning mode—consume the opening tag at the prompt level, so the generated output begins directly with the reasoning content and terminates with a bare `</think>` rather than a balanced tag pair. To handle this case, the parser additionally discards all content preceding and including any residual `</think>` token, regardless of whether a matching opening tag is present.

Step 2 — Markdown Fence Removal Although the system prompt explicitly instructs models to return a JSON array with no markdown formatting, some models nonetheless wrap their output in code fences (‘`json ... `‘), a known failure mode of instruction-tuned models on structured generation tasks. Code fences are removed unconditionally before the string is submitted to the JSON parser.

Step 3 — JSON Parsing and Type Validation The cleaned string is submitted to a strict JSON parser. A failure at this stage—whether due to malformed syntax, incomplete generation, or stripping-resistant natural-language preamble—is recorded as `parse_success = False`; the sample contributes an empty prediction array to all downstream entity-level metrics, injecting false negatives for every ground-truth entity on the page. Reporting the Format Compliance metric (defined in §B.7) disentangles these format-induced false negatives from genuine reasoning failures. A successful parse is additionally type-checked: a response that parses as a JSON object rather than an array is rejected and treated as a parse failure, since the output contract specified in §4.3 requires an array. On parse failure, up to three retries are issued for the same sample before recording a permanent failure; format compliance of 100% is achieved for all evaluated models within this retry budget.

Step 4 — Coordinate Rescaling As described in §4.1, models are prompted to emit bounding-box coordinates as integers in $[0, 1000]$ to reduce output token count and eliminate floating-point formatting variance across inference backends. After parsing, coordinates are rescaled to $[0, 1]$ by dividing by $1{,}000$ when the maximum observed coordinate value in the response exceeds 1.0; this heuristic is safe by construction because valid nor-

Prompt template — inference (all models)

System message:

You are a medical document de-identification expert.
Identify all Protected Health Information (PHI) in the document page image.

For each PHI entity return:
`bbox_2d` – list of one or more `[y_min, x_min, y_max, x_max]` bounding boxes, one per line of text (e.g. a name split across two lines gets two boxes). Coordinates are integers in `[0, 1000]`, normalised relative to page height (y) and page width (x).

`label` – one of the following:
`{labels_with_guidelines}`

`text` – exact text string as it appears in the image.

Return ONLY a valid JSON array. No explanation, no markdown fences.
If no PHI is present return: `[]`

Example output:

```
[
  {"bbox_2d": [[120, 50, 140, 280]], "label": "NAME:PATIENT", "text": "Mario Rossi"},
  {"bbox_2d": [[220, 100, 240, 350]], "label": "DATETIME", "text": "15/03/2024"}
]
```

User message:

`<page_image>`

`List all PHI entities in this document page.`

Figure 10: **Inference prompt template submitted to all evaluated models.** The `{labels_with_guidelines}` placeholder is expanded at runtime with the complete PHI label set and corresponding annotation guidelines. The user message provides the document page image directly as multimodal input, represented here with the symbolic placeholder `<page_image>`.

malized coordinates are bounded in $[0, 1]$, and the two regimes are non-overlapping for any coordinate above 1. All metric computations operate exclusively on $[0, 1]$-normalized coordinates.

We achieve complete format adherence within 3 retries from the model during evaluation but we also tested guided decoding through vLLM,[15] enforcing structural compliance via the xgrammar[16] constrained generation engine (Willard and Louf, 2023). This kind of configuration leads to worse overall performance. Localization decreases significantly, and in some instances infinite loops lead to truncated outputs, since it forces models to go out of distribution, as already highlighted by Beurer-Kellner et al. (2024).

B.5 Hardware and Model Serving Infrastructure

Open-weight models are served locally via vLLM (Kwon et al., 2023), whose PagedAttention memory manager enables efficient KV-cache utilisation at the long generation budgets required by reasoning-capable models. Each model is launched inside a Docker container and exposed on a configurable port as an OpenAI-compatible REST endpoint; the inference script connects over localhost with no external network routing, keeping inference local (Cerroni et al., 2013, 2015). All experiments are conducted on a cluster of 4 NVIDIA H100 80 GB GPUs.

B.6 GLiNER Label Mapping

GLiNER-based models are *open-domain*: rather than relying on a fixed set of entity types, they are conditioned at inference time on natural-language labels. This allows them to be queried with arbitrary label definitions. To align the models with the 23 categories in our HIPAA-based PIB taxonomy, we therefore mapped each category to a corresponding natural-language label, as shown in Table 13. We acknowledge the existence of other zero-shot encoder-based NER models, such as OpenBioNER (Cocchieri et al., 2025). However,

[15] https://docs.vllm.ai/en/v0.8.2/features/structured_outputs.html

[16] https://github.com/mlc-ai/xgramar

Category	Label	Description
Name	NAME:PATIENT	Patient name; honorifics excluded; partial identifiers (isolated surname, period-separated initials) included.
	NAME:STAFF	Medical staff name (physicians, nurses, technicians); titles excluded; partial identifiers included.
	NAME:ASSOCIATE	Name of a non-staff associate of the patient (family member, caregiver); honorifics excluded.
	NAME:FACILITY	Name of a hospital, clinic, laboratory, or healthcare facility.
	NAME:DEPARTMENT	Name of a department, ward, or unit within a facility.
Date of Birth	DATE_OF_BIRTH:UNDER_89	Date of birth of a patient aged 89 or younger.
	DATE_OF_BIRTH:OVER_89†	Date of birth of a patient aged over 89; directly identifying under HIPAA Safe Harbor.
Datetime	DATETIME	Any date or date-time other than date of birth, including isolated years in narrative prose.
Age	AGE:UNDER_89	Explicit age of a patient 89 years old or younger.
	AGE:OVER_89†	Explicit age of a patient over 89; directly identifying under HIPAA Safe Harbor.
ID	ID:PATIENT_ID	Static per-patient identifier (e.g. medical record number).
	ID:DOCUMENT_ID	Document or report identifier (e.g. transcription code).
	ID:SPECIMEN_ID	Specimen or biological sample identifier.
	ID:STAFF_ID	Staff identifier (badge number, employee ID, licence number).
	ID:DEVICE_ID	Medical device identifier or serial number.
	ID:EXAM_ID	Examination or procedure identifier.
	ID:ADMISSION_ID	Episode-specific hospitalisation or encounter identifier.
Contact	CONTACT:PATIENT	Phone, fax, or e-mail of the patient.
	CONTACT:STAFF	Personal contact explicitly attributed to a specific staff member; use only when clearly personal, not institutional.
	CONTACT:ASSOCIATE	Contact information of a non-staff associate of the patient.
	CONTACT:FACILITY	Phone, fax, or e-mail of the facility or department; facility and department numbers are merged into this category.
Address	ADDRESS:PATIENT	Geographic address of the patient, from street level to city and country.
	ADDRESS:FACILITY	Geographic address of a healthcare facility, from street level to city and country.

Table 12: **PIB label taxonomy.** All 23 entity types derived from the HIPAA Safe Harbor specification, grouped into six macro-categories, along with the guidelines used at inference time. † = directly identifying under Safe Harbor regardless of age, requiring a distinct redaction regime.

we do not include them in this evaluation because they require richer entity specifications, such as textual descriptions, rather than only natural-language labels. Including such models would therefore introduce additional prompt and label-engineering choices that are not directly comparable to the standardized label-mapping protocol used for GLiNER-based models. We omit these models to maintain a consistent and controlled evaluation setting across the evaluated GLiNER-based approaches.

B.7 Metrics Formulation

This section formalizes the complete metric suite evaluated on PIB, extending the definitions introduced in §4.4 with the additional diagnostic metrics used in complementary experiments discussed in § B.8, § B.10, § B.11. Throughout, a PHI annotation record—whether a model prediction or a ground-truth instance—is represented as a triple $(\ell,\, t,\, B)$: a label ℓ drawn from the 23-type taxonomy, a verbatim text string t, and a set of axis-aligned bounding boxes $B = \{b_1, \ldots, b_k\}$ with $k \geq 1$, each expressed in $[0, 1]$-normalized coordinates $[y_{\min}, x_{\min}, y_{\max}, x_{\max}]$ relative to page height and width respectively. When an entity is annotated with $k > 1$ disjoint bounding boxes (e.g. a patient address split across two typeset lines), computing IoU over the union of polygon fragments is geometrically complex. It requires evaluating the area of the intersection and union of two arbitrary sets of rectangles, which may themselves overlap within each set, necessitating a computational geometry routine to avoid double-counting interior regions. We thus reduce all multi-box spans

PIB Label	Natural-language label
NAME:PATIENT	patient name
NAME:STAFF	healthcare staff name
NAME:ASSOCIATE	patient associate name
NAME:FACILITY	healthcare facility name
NAME:DEPARTMENT	hospital department name
DATE_OF_BIRTH:UNDER_89	date of birth
DATE_OF_BIRTH:OVER_89	date of birth of elderly patient
DATETIME	date or time value
AGE:UNDER_89	patient age
AGE:OVER_89	age of elderly patient
ID:PATIENT_ID	patient identifier
ID:DOCUMENT_ID	document identifier
ID:SPECIMEN_ID	specimen identifier
ID:STAFF_ID	staff identifier
ID:DEVICE_ID	device identifier
ID:EXAM_ID	examination identifier
ID:ADMISSION_ID	admission identifier
CONTACT:PATIENT	patient contact information
CONTACT:STAFF	staff contact information
CONTACT:ASSOCIATE	associate contact information
CONTACT:FACILITY	facility contact information
ADDRESS:PATIENT	patient address
ADDRESS:FACILITY	facility address

Table 13: Mapping between PIB taxonomy labels and the natural-language labels used to query GLiNER-based models.

to the smallest enclosing rectangle, where index $i \in \{1, \ldots, k\}$ ranges over the individual boxes in B:

$$\begin{aligned}\mathrm{enc}(B) = \big[&\min_i y^i_{\min},\ \min_i x^i_{\min}, \\ &\max_i y^i_{\max},\ \max_i x^i_{\max} \big]\end{aligned} \tag{1}$$

Because $\mathrm{enc}(B)$ stretches from the topmost to the bottommost fragment, it absorbs the inter-fragment whitespace between line breaks into its area. This inflates the IoU denominator (union area) without a corresponding increase in the numerator (intersection area), systematically suppressing IoU scores for fragmented spans even when a model localizes each fragment correctly. The bias is most pronounced for ADDRESS:PATIENT (31.0% multi-bbox rate) and CONTACT:ASSOCIATE (22.2%), and should be borne in mind when interpreting per-label spatial results for these categories.

Greedy Bipartite Matching All metrics employ greedy bipartite matching (Sang and Meulder, 2003): candidate prediction–ground-truth pairs sharing the same label ℓ are scored, sorted in descending order, and accepted greedily, with each entity matched at most once. Predictions with no same-label ground-truth counterpart count directly as false positives; ground-truth entities with no same-label prediction count as false negatives.

The following four metrics are reported as primary scalars in §5.

Exact Span F1 Let $\mathrm{FP}_{\mathrm{ex}}$ denote predictions with no exact-match ground-truth counterpart, and $\mathrm{FN}_{\mathrm{ex}}$ ground-truth entities for which no exact-match prediction exists:

$$P_{\mathrm{ex}} = \frac{|\mathcal{M}_{\mathrm{ex}}|}{|\mathcal{M}_{\mathrm{ex}}| + \mathrm{FP}_{\mathrm{ex}}},\ R_{\mathrm{ex}} = \frac{|\mathcal{M}_{\mathrm{ex}}|}{|\mathcal{M}_{\mathrm{ex}}| + \mathrm{FN}_{\mathrm{ex}}} \tag{2}$$

$$F_{\mathrm{ex}} = \frac{2\,P_{\mathrm{ex}}\,R_{\mathrm{ex}}}{P_{\mathrm{ex}} + R_{\mathrm{ex}}} \tag{3}$$

where $\mathcal{M}_{\mathrm{ex}}$ denotes the set of exact-match pairs. Because a near-correct transcription (e.g., a single character substitution) simultaneously increments both $\mathrm{FP}_{\mathrm{ex}}$ and $\mathrm{FN}_{\mathrm{ex}}$, this metric applies strictly more pressure than partial-credit variants.

Spatial F1 At each threshold τ, a prediction is a true positive if it shares the same label ℓ with its matched ground-truth entity and $\mathrm{IoU}(\mathrm{enc}(B_p), \mathrm{enc}(B_g)) \geq \tau$, where B_p and B_g are the bounding-box sets of the predicted and ground-truth entity respectively. $\mathrm{F1}_{@\tau}$ is then the macro-averaged F1 over entity types at that threshold. Spatial F1 averages across three thresholds, following the COCO mAP convention (Everingham et al., 2010; Lin et al., 2014):

$$\begin{aligned}\text{Spatial-F1} &= \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \mathrm{F1}_{@\tau}, \\ \mathcal{T} &= \{0.25,\ 0.50,\ 0.75\}\end{aligned} \tag{4}$$

Entity Mean IoU For every ground-truth entity g with label ℓ, the highest IoU against any same-label prediction $p \in \hat{P}_\ell$ is recorded independently, where $\hat{P}_\ell$ denotes the set of all predictions with label ℓ; unlike the greedy-matched metrics, a single prediction may contribute to the score of multiple ground-truth entities.

$$\begin{aligned}\mathrm{IoU}_{\mathrm{EM}} = \frac{1}{|GT|} \sum_{g \in GT} &\max_{p \in \hat{P}_{g.\ell}} \\ &\mathrm{IoU}(\mathrm{enc}(B_p),\ \mathrm{enc}(B_g))\end{aligned} \tag{5}$$

with max over an empty set defined as 0. Missed entities contribute 0 and depress the numerator without being excluded from the denominator. The deliberate departure from greedy matching serves a diagnostic purpose: $\mathrm{IoU}_{\mathrm{EM}}$ measures the *best achievable* spatial coverage for each ground-truth entity, decoupled from whether the model also emits spurious predictions elsewhere. A model

that correctly localizes every PHI token but surrounds each box with redundant duplicates would score high on $\mathrm{IoU}_{\mathrm{EM}}$ yet low on Spatial F1, whose precision term penalizes exactly those false positives; the divergence between the two metrics therefore directly quantifies the redundancy cost of over-prediction, a failure mode distinct from both mislocalization and missed detection.

Pass Rate Let $\mathrm{TP}_{\mathrm{bbox}}$ denote the set of prediction–ground-truth pairs accepted as true positives under the spatial criterion ($\mathrm{IoU} > 0.5$, same label), and let $p[t].\mathrm{norm}$ and $g[t].\mathrm{norm}$ denote the lowercased, whitespace-stripped surface forms of the predicted and ground-truth text strings respectively. Pass Rate is the fraction of spatially matched pairs whose transcriptions also agree exactly:

$$\mathrm{PR} = \frac{1}{|\mathrm{TP}_{\mathrm{bbox}}|}\Big|\{(p,g) \in \mathrm{TP}_{\mathrm{bbox}} : p[t].\mathrm{norm} = g[t].\mathrm{norm}\}\Big| \quad (6)$$

Jointly requiring correct label, $\mathrm{IoU}{>}0.5$, and exact surface-form reproduction, this is the strictest scalar in the suite.

The following metrics are computed for ablation analyses.

Approximate Span F1 A relaxed variant of Exact Span F1 in which a prediction is accepted as a true positive when its character-level F1 against the ground-truth string exceeds 0.5 rather than requiring exact match. This threshold corresponds to the minimum charF1 achievable by any non-empty common subsequence, tolerating minor transcription artifacts without crediting substantially wrong or hallucinated spans. Comparing Approximate against Exact Span F1 isolates the marginal cost of requiring verbatim surface-form reproduction.

Character-level F1 Let $\mathrm{TP}_{\mathrm{text}}$ denote the set of prediction–ground-truth pairs accepted as true positives under the Approximate Span F1 criterion ($\mathrm{charF1} > 0.5$, same label), and let $\mathrm{charF1}(p[t], g[t])$ denote the character-level F1 between the predicted and ground-truth surface forms, computed as the harmonic mean of character-precision and character-recall over their longest common character subsequence. Mean charF1 is then averaged exclusively over $\mathrm{TP}_{\mathrm{text}}$:

$$\mathrm{charF1}_{\mathrm{text}} = \frac{1}{|\mathrm{TP}_{\mathrm{text}}|} \sum_{(p,g) \in \mathrm{TP}_{\mathrm{text}}} \mathrm{charF1}(p[t],\, g[t]) \quad (7)$$

Since $\mathrm{TP}_{\mathrm{text}}$ already gates on $\mathrm{charF1}{>}0.5$, this metric is conditioned on matched pairs and measures transcription *fidelity* independently of detection performance. Values near 0.5 flag systematic borderline transcriptions, while values near 1.0 confirm near-perfect surface-form reproduction on matched spans. Its complement $1 - \mathrm{charF1}_{\mathrm{text}}$ quantifies the residual character-level imprecision that the Approximate Span F1 threshold tolerates but Exact Span F1 penalizes.

Format Compliance Let N denote the total number of evaluated page instances and $\mathrm{parse\_success}_i \in \{\texttt{True}, \texttt{False}\}$ indicate whether the model response for instance i survives the full normalization pipeline of §B.4 (chain-of-thought stripping, markdown fence removal, JSON parsing, and type validation) without error:

$$\mathrm{FC} = \frac{\big|\{i : \mathrm{parse\_success}_i = \texttt{True}\}\big|}{N} \quad (8)$$

FC measures output format adherence without constrained decoding rather than PHI extraction capability. It is computed before any metric to prevent parse failures—which silently inject empty prediction arrays, inflating false negatives in a way indistinguishable from genuine reasoning failures—from contaminating the evaluation surface.

B.8 Frequency-Weighted Exact Span F1

Micro averaging weights each entity occurrence equally, so common entity types dominate the score (Table 14). Gemma-4-31B leads at 81.6 average micro F1, with Qwen3.5-27B (78.1) and Qwen3.5-122B-A10B (77.2) close behind; the smaller gap between these two Qwen variants here—compared to ApxF1—suggests the MoE model is weaker on frequent entity types. InternVL3.5-14B and Qwen3.6-35B both exhibit a pronounced precision–recall gap (precision ≈ 65–81, recall ≈ 37–64), confirming systematic under-prediction of span boundaries across the most common entity types. Qwen3.5-9B shows the steepest degradation from Base to Hard (-8.4 micro F1), suggesting its extraction of common entities is particularly sensitive to document complexity. MedGemma-27B-it collapses to 3.3 average, consistent with its inability to produce coherent entity spans regardless of frequency.

B.9 Per-Label Exact Span F1

Table 15 reports span-exact F1 for each of the 23 entity types. The clearest predictor of per-

Model	Base			Medium			Hard			AVG
	F1	P	R	F1	P	R	F1	P	R	F1
Gemma-4-31B-it	**82.8**	**84.4**	**81.2**	**82.6**	**84.3**	**81.0**	**79.5**	**81.5**	**77.7**	**81.6**
Qwen3.5-27B	78.8	81.2	76.6	79.2	78.7	79.7	76.2	76.1	76.3	78.1
Gemma-4-26B-A4B-it	78.9	80.8	77.0	79.1	80.6	77.5	74.7	76.2	73.2	77.5
Qwen3.5-122B-A10B	79.4	79.3	79.4	77.1	77.0	77.2	75.2	75.4	75.0	77.2
Qwen3.5-9B	76.4	79.1	73.9	72.6	73.6	71.6	68.0	71.3	65.0	72.3
Qwen3.6-35B-A3B	71.7	81.4	64.0	72.6	80.9	65.9	69.5	78.8	62.1	71.3
InternVL3.5-14B	52.0	64.9	43.4	48.6	60.5	40.6	45.1	57.6	37.1	48.6
MedGemma-27B-it	3.3	2.9	3.7	3.0	3.1	2.9	3.7	4.5	3.2	3.3

Table 14: **Exact Span F1 with micro averaging on PIB.** Reported as a frequency-weighted complement to Table 2, where all 23 label types contribute equally. Scores are uniformly higher than their macro counterparts, as common administrative entities are easier to extract; model rankings are broadly preserved. Models are sorted by decreasing average F1. Best scores in **bold**, second-best underlined.

formance is *lexical regularity*: date-of-birth entities (DATE_OF_BIRTH:UNDER_89) are the highest-scoring label across the board, with top models reaching 97.6 F1, as their surface form is almost invariant. At the other extreme, ID:DEVICE_ID scores as low as 11.8 (Qwen3.5-9B) and no higher than 50.0 even for the best model, reflecting the complete absence of a canonical format for medical device serial numbers. Within the NAME group, NAME:ASSOCIATE is unexpectedly the easiest sub-type (83.4), likely because non-staff associates are always explicitly introduced in context, whereas NAME:DEPARTMENT is the hardest (50.7), as department names are frequently abbreviated or embedded in longer institutional strings. A consistent asymmetry appears in both CONTACT and ADDRESS: models recover patient contacts and addresses substantially less reliably than facility-level counterparts (51.9 vs. 75.3 F1 for CONTACT:STAFF/PATIENT; 51.8 vs. 65.8 for ADDRESS:PATIENT/FACILITY), pointing to greater surface variability in patient-facing fields. The ID group shows the widest intra-category spread of any group ($\Delta = 48$ points between ID:DOCUMENT_ID and ID:DEVICE_ID), underscoring that identifier difficulty is driven not by the category but by the standardization of individual identifier types.

B.10 Transcription Fidelity on Matched Entities

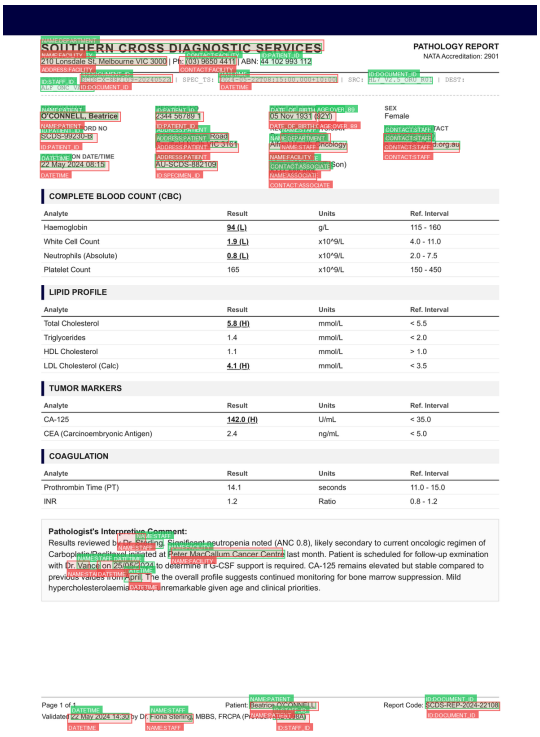
CharF1 is uniformly high for all well-formed models (97–99%), confirming that near-correct transcription is largely independent of entity recall (Table 16). The ∼10-point drop from CharF1 to EMR for top models reflects residual casing, whitespace, or minor boundary differences that prevent exact string matches. Comparing EMR and ApxF1 exposes two distinct failure modes: Qwen3.6-35B scores high EMR (88.1 on Base) but low ApxF1 (72.9), indicating it finds fewer entities but transcribes them accurately; InternVL3.5-14B shows the opposite pattern (EMR 70.3, ApxF1 74.1), finding more entities but with lower verbatim precision. MedGemma-27B-it is the clear outlier: CharF1$\approx$ 78 but EMR$\approx$ 18, consistent with a model that produces plausible but malformed output rather than failing at character-level transcription. All three metrics degrade modestly but consistently from Base to Hard (typically 3–7 EMR points), with CharF1 proving the most robust (< 2 points of degradation for top models).

B.11 Spatial Grounding Across IoU Thresholds

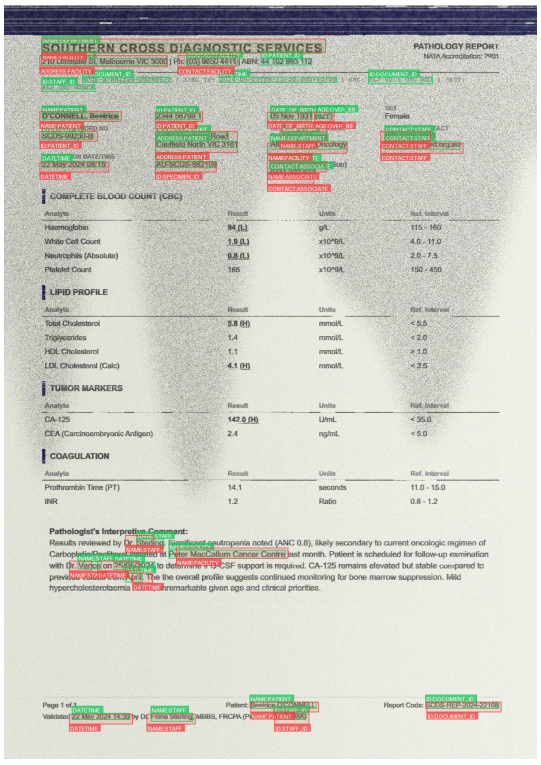
Tightening the IoU threshold from τ=0.25 to τ=0.75 causes near-total collapse across all models, revealing that bounding boxes are at best approximately correct (Table 17). At τ=0.25, Gemma-4-31B ($\approx$ 75.6 on Base) and Gemma-4-26B-A4B ($\approx$ 72.5) maintain competitive scores, while Qwen3.5-27B lags at $\approx$ 60–65 and all other models are essentially zero. By τ=0.75, even Gemma-4-31B drops to $\approx$ 17–20 F1, while every Qwen variant, InternVL, and MedGemma score below 3—confirming that only the dense Gemma family produces boxes tight enough to survive strict spatial evaluation. Scores are stable across Base / Medium / Hard within each model, implying that difficulty affects which entities are found, not how precisely the found ones are localised.

B.12 Result Samples

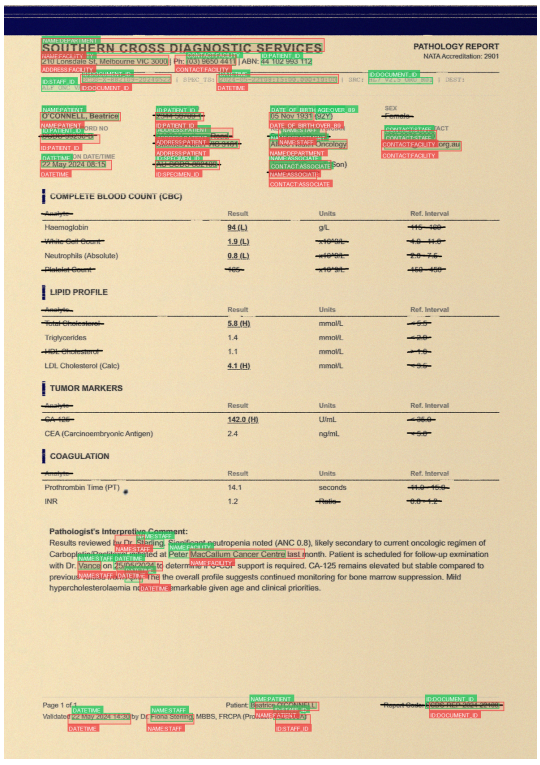
Fig. 11 illustrates how extraction quality for a single document degrades across splits for Gemma-4-31B-it, the strongest open-weight model evalu-



(a) **Base split.** Clean 144 DPI rasterization.



(b) **Medium split.** ColorPaper tint, LightingGradient, and InkBleed.



(c) **Hard split.** Faxify binarization, InkBleed, Lines-Degradation, and LowInk banding.

Figure 11: **Qualitative example: Gemma-4-31B-it across difficulty splits.** Model predictions (red) against ground truth (green) for the same source document (`sample_0005_lab_australia_female.pdf`, page 1) at Base, Medium, and Hard. Bounding boxes and PHI labels are annotated once on the clean page and held fixed across splits, so any difference between panels is attributable to the model rather than to the annotation.

	Date	Name						DoB			Age			ID								Contact					Address		
Model	DT	Pat.	Staff	Assoc.	Fac.	Dept.	*Avg*	<89	>89	*Avg*	<89	>89	*Avg*	Pat.	Doc.	Spec.	Staff	Dev.	Exam	Adm.	*Avg*	Pat.	Staff	Assoc.	Fac.	*Avg*	Pat.	Fac.	*Avg*
Gemma-4-31B	89.1	**84.9**	**94.8**	**92.8**	**73.1**	55.4	**80.2**	**97.6**	**94.9**	**96.2**	66.5	**79.6**	73.0	77.8	**81.1**	66.7	**74.4**	28.6	46.0	61.0	62.2	**82.9**	50.9	**88.2**	71.8	**73.5**	**65.4**	**74.8**	**70.1**
Gemma-4-26B-A4B	86.1	81.6	92.2	88.7	65.0	47.5	75.0	95.6	91.6	93.6	60.1	69.4	64.8	76.1	73.3	**71.8**	74.1	14.3	41.9	59.6	58.7	81.8	53.3	67.5	**76.0**	69.6	50.8	70.6	60.7
Qwen3.5-27B	**89.2**	82.5	86.7	88.8	68.3	58.8	77.0	97.5	88.8	93.2	**80.3**	77.5	**78.9**	**81.0**	74.8	68.9	72.7	**50.0**	54.3	61.4	**66.2**	80.0	**65.4**	73.5	72.5	72.8	51.5	69.5	60.5
Qwen3.6-35B-A3B	78.6	73.4	81.2	77.7	65.5	51.8	69.9	86.1	77.5	81.8	67.2	68.3	67.8	69.0	71.9	54.5	67.8	15.4	**61.5**	51.5	56.0	73.4	50.0	72.6	61.7	64.4	46.1	60.2	53.1
Qwen3.5-122B-A10B	88.3	83.0	79.4	83.1	69.0	**60.4**	75.0	**97.6**	89.7	93.7	69.2	82.5	75.9	79.3	76.9	71.1	72.6	26.7	58.1	**62.0**	63.8	78.7	59.6	65.2	68.2	67.9	51.3	70.7	61.0
Qwen3.5-9B	82.1	77.1	81.3	76.0	64.4	48.9	69.5	73.7	71.9	72.8	62.8	67.2	65.0	71.4	73.2	62.3	64.1	11.8	56.9	57.3	56.7	73.9	44.0	82.3	63.7	66.0	52.1	64.3	58.2
InternVL3.5-14B	51.8	56.7	46.6	76.8	45.4	32.0	51.5	81.6	74.6	78.1	45.6	51.1	48.3	57.9	40.6	49.3	55.0	28.6	33.3	51.1	45.1	56.1	40.0	67.9	17.4	45.3	45.5	50.8	48.2
MedGemma-27B-it	5.4	7.7	2.4	5.0	1.5	0.0	3.3	11.7	4.8	8.2	1.7	0.0	0.8	2.4	0.8	0.0	0.0	0.0	0.0	0.0	0.5	1.7	0.0	1.9	0.7	1.1	4.1	0.5	2.3

Table 15: **Per-label Exact Span F1 on PIB.** Results are reported for each of the 23 entity types in the taxonomy, grouped by macro-category (Datetime / Name / Date of Birth / Age / ID / Contact / Address) and aggregated across all difficulty splits and document domains. *Avg* denotes the macro-average F1 within each group (omitted for single-label groups). Models are sorted in the same order as Table 2. Best scores are reported in **bold**, while second-best scores are underlined.

	Base			Medium			Hard			AVG
Model	CharF1	EMR	ApxF1	CharF1	EMR	ApxF1	CharF1	EMR	ApxF1	EMR
Gemma-4-31B-it	**98.8**	**89.3**	85.2	**98.8**	**89.5**	84.8	**98.3**	**86.4**	81.2	**88.4**
Gemma-4-26B-A4B-it	98.5	88.2	78.5	98.4	88.1	81.9	97.6	83.3	82.6	86.5
Qwen3.6-35B-A3B	98.2	88.1	72.9	98.2	87.3	77.4	97.7	84.2	73.9	86.5
Qwen3.5-27B-FP8	98.2	87.6	83.0	97.9	85.9	**85.4**	97.5	83.0	**86.5**	85.5
Qwen3.5-9B	98.4	87.5	78.6	97.9	85.0	78.5	97.2	81.1	75.7	84.5
Qwen3.5-122B-A10B-FP8	98.0	86.0	**86.2**	97.7	83.5	**85.4**	97.4	81.7	85.6	83.7
InternVL3-5-14B	95.7	70.3	74.1	95.0	66.0	73.0	94.2	62.5	70.3	66.3
MedGemma-27b-it	78.4	18.6	10.9	77.4	16.2	11.1	77.3	19.5	11.9	18.1

Table 16: **Transcription fidelity on text-matched entities. CharF1**: SQuAD-style character-level F1 averaged over text-matched entity pairs. **EMR**: fraction of text-matched pairs with exact case-insensitive string match. **ApxF1**: approximate span F1 (detection-level, char-F1 > 0.5 threshold). Models are sorted by decreasing average EMR. Best scores in **bold**, second-best underlined.

	Base			Medium			Hard		
Model	τ=.25	τ=.50	τ=.75	τ=.25	τ=.50	τ=.75	τ=.25	τ=.50	τ=.75
Gemma-4-31B-it	**75.6**	**64.6**	**20.5**	**75.8**	**64.4**	**18.0**	**75.5**	**62.8**	**15.9**
Gemma-4-26B-A4B-it	72.5	58.3	14.4	75.0	58.0	15.4	74.5	54.1	14.4
Qwen3.5-27B	60.6	40.0	13.1	64.2	42.8	12.1	65.0	42.9	12.9
Qwen3.6-35B-A3B	30.5	13.0	1.1	41.4	20.5	2.6	33.2	15.1	1.7
Qwen3.5-122B-A10B	25.0	8.1	0.4	34.1	15.1	1.8	32.5	13.5	2.7
Qwen3.5-9B	2.4	0.5	0.0	5.3	0.8	0.0	6.1	1.5	0.1
InternVL3.5-14B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MedGemma-27B-it	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 17: **Spatial F1 across IoU thresholds on PIB.** Results are reported at $\tau \in \{0.25, 0.50, 0.75\}$ (macro-averaged over entity types) across Base / Medium / Hard splits. Models are sorted by decreasing overall average PR (see Table 2). Best scores in **bold**, second-best underlined.

ated. On the Base split (Fig. 11a), the model recovers 28 of the 29 ground-truth entities exactly, with the sole error being a transposed postcode in `ADDRESS:PATIENT`. On the Hard split (Fig. 11c), performance does not degrade through a simple loss of recall. The model still emits 28 entities, and their bounding boxes remain close to the ground-truth regions. This is consistent with the aggregate finding that Spatial F1 falls by only 1.6 points from Base to Hard, compared to an 8.4-point drop in Text F1 (Table 2). Instead, degradation occurs in the text generated inside those boxes, manifesting through three main failure modes. First, one entity is omitted outright: the `AGE:OVER_89` span (`92Y`), correctly extracted on Base and Medium, is entirely absent from the Hard output. This is the most consequential loss on the page, since ages above 89 are directly identifying under HIPAA Safe Harbor rules and carry the same redaction priority as a full name. Second, and most frequently, six identifier and contact spans are detected under the correct label and location, but suffer from transcription corruption—digits are substituted within patient, specimen, and document identifiers; hyphens are flattened to spaces; and a street number is altered. Because each corrupted value preserves the original field format and character count, the model is confidently wrong rather than abstaining. This is a par-

ticularly dangerous outcome for a de-identification system, as a mistranscribed value leaves a residual, altered identifier in the output. Third, the two `CONTACT:STAFF` entities belonging to the reporting oncologist are relabelled as `CONTACT:FACILITY` on Hard, though their text remains unaffected. While degradation is not strictly monotonic at the span level—one facility name is mislabelled on Base and Medium but assigned correctly on Hard—the overall pattern is clear. Harder documents do not merely cause the model to miss PHI; they also corrupt the PHI it does attempt to redact.

B.13 MedGemma: Failure Mode Analysis

MedGemma-27B achieves an average Exact Span F1 of 2.3% across the three difficulty splits, the lowest in the cohort by a margin of 47.8 points. Crucially, this collapse is *not* attributable to output structure: Format Compliance is 100%, confirming the model consistently produces well-formed JSON arrays and that all scores are free from format-induced artifacts.

The failure instead decomposes into two simultaneous pathologies. First, the model generates a large volume of plausible-looking but spurious entities, yielding hallucination rates of 77–84% ($\mathrm{FP}/(\mathrm{TP}+\mathrm{FP})$, text-based) across splits. Second, it simultaneously misses 80–84% of ground-truth PHI instances. This joint pattern of high false-positive and high-false-negative rates is consistent with a model whose entity generation is governed by prior distributional knowledge of what *kind* of tokens are medically plausible, rather than by exhaustive extraction of what is actually *present* on the page.

A principled account of this behavior follows from inspecting MedGemma's training distribution. The MedSigLIP vision encoder powering the model was pre-trained on over 33M medical image–text pairs drawn exclusively from clinical imaging modalities: chest X-rays, histopathology patches, dermatology images, ophthalmology images, and CT/MRI slices. None of these are text-dense document images. Any spatial grounding capability acquired during training targets *anatomical localization of imaging findings*—a fundamentally different task from pixel-level NER over typeset PDF layouts, where the visual signal is character strokes and tabular geometry rather than pathological tissue patterns. PIB thus lies entirely outside the model's visual pre-training distribution, and its SigLIP encoder cannot be expected to produce spatially discriminative features for text-region localization.

PIB Annotation Guidelines

Your task. You will be shown one page of a clinical PDF document at a time. For every piece of Protected Health Information (PHI) visible on the page, draw a tight bounding box around it and assign the correct label from the taxonomy below. You must also transcribe the exact text string as it appears in the image.
What to annotate. Annotate all tokens that fall under the 23-category HIPAA-aligned taxonomy (see label list below). This includes identifiers embedded in narrative prose (e.g., a hospital name mentioned mid-sentence, an isolated year in a referral paragraph, or a patient surname appearing only in the page footer).
What NOT to annotate. Do *not* annotate non-identifying clinical content: diagnoses, medications, lab values, imaging findings, procedure names, anatomical terms, or clinical abbreviations that do not identify a person, facility, or device.

Bounding boxes.

- Draw the box as tightly as possible around the token(s), excluding surrounding whitespace and punctuation that is not part of the identifier (e.g., exclude a trailing comma or a parenthetical period).
- If an entity wraps across a line break, draw *two separate* disjoint boxes—one per line fragment—and attach both to the same annotation record using multi-box mode.
- Use the sub-pixel nudge controls (± 1, ± 5 px) to achieve tight alignment, especially for entities evaluated at the strict IoU ≥ 0.75 threshold.
- For spatially regular fields (dates, patient IDs), the fixed-size stamp mode centres a preset box on your click point and speeds up annotation without sacrificing precision.

Text transcription. Copy the surface form *verbatim* as printed in the image, preserving capitalisation, punctuation internal to the token, and spacing. Do not expand abbreviations or normalise formats (e.g., transcribe "07/12/1951" as-is, not as "December 7, 1951").

Label taxonomy and disambiguation rules.

Label	What to include / exclude
`NAME:PATIENT`	Patient name. Exclude honorifics (Mr., Mrs.); include isolated surnames and period-separated initials (e.g., M.R.).
`NAME:STAFF`	Physician, nurse, or technician name. Exclude titles (Dr., Prof.); include partial identifiers.
`NAME:ASSOCIATE`	Non-staff person associated with the patient (family member, caregiver). Exclude honorifics.
`NAME:FACILITY`	Hospital, clinic, laboratory, or healthcare facility name.
`NAME:DEPARTMENT`	Department, ward, or unit within a facility.
`DATE_OF_BIRTH:UNDER_89`	Date of birth of a patient aged 89 or younger.
`DATE_OF_BIRTH:OVER_89`	Date of birth of a patient aged over 89. Directly identifying under HIPAA Safe Harbor—treat with the same priority as a full name.
`DATETIME`	Any date or date-time *other than* date of birth, including isolated years in narrative prose (e.g., "a procedure performed in 2019").
`AGE:UNDER_89`	Explicit numeric age of a patient 89 years old or younger.
`AGE:OVER_89`	Explicit numeric age of a patient over 89. Directly identifying under Safe Harbor.
`ID:PATIENT_ID`	Static per-patient identifier (medical record number, NHS number, SSN, etc.).
`ID:DOCUMENT_ID`	Report or transcription code.
`ID:SPECIMEN_ID`	Specimen or biological sample barcode / identifier.
`ID:STAFF_ID`	Badge number, employee ID, or medical licence number of a staff member.
`ID:DEVICE_ID`	Medical device serial number or equipment identifier.
`ID:EXAM_ID`	Examination or accession number.
`ID:ADMISSION_ID`	Episode-specific hospitalisation or encounter number.
`CONTACT:PATIENT`	Phone, fax, or e-mail of the patient.
`CONTACT:STAFF`	Personal contact of a specific staff member (e.g., a doctor's direct mobile number). Use *only* when the contact is clearly personal, not institutional.
`CONTACT:ASSOCIATE`	Contact information of a non-staff associate.
`CONTACT:FACILITY`	Phone, fax, or e-mail of the facility or a department. Facility and department numbers are merged into this single label.
`ADDRESS:PATIENT`	Geographic address of the patient, from street level to city and country.
`ADDRESS:FACILITY`	Geographic address of a healthcare facility, from street level to city and country.

Common edge cases.

- *Partial names*: isolated surnames, first initials, or period-separated initials (M.R.) are PHI and must be annotated.
- *Contextual PHI*: a facility name or physician name that appears only once in a mid-sentence reference still requires annotation.
- *Ambiguous staff contact*: a phone number in the letterhead labelled "Department of Radiology" goes to `CONTACT:FACILITY`; a number explicitly attributed to "Dr. Smith (direct line)" goes to `CONTACT:STAFF`.
- *Age vs. date of birth*: annotate the token that is actually present. If both appear, annotate each separately with its own label.
- *Non-PHI clinical numbers*: lab result values, reference ranges, medication doses, imaging measurements, and procedure codes are **not** PHI and must not be annotated.

Figure 12: **PIB human annotation guidelines.** Distributed to all annotators before labeling and kept open as a reference throughout. Disagreements between the two independent annotators were resolved by adjudication against these rules.