

The Diagnosis Is Already in the Notes: A Channel-Wise Audit of Diagnostic Shortcuts in Clinical-Text Benchmarks

Haocun Ye^{1,2}, Xinlong Jiang^{1,2†}, Yuanyuan Hu³, Shubai Chen^{1,2}, Tingyu Wu^{1,2},
Zhenkun Zheng^{1,2}, Caijiao Yi³, Teng Zhang^{1,2}, Liuxin Wang^{1,2}, Tong Zhao^{1,2},
Heping Xu³, Weiwei Dai³, Yiqiang Chen^{1,2}

¹Beijing Key Laboratory of Mobile Computing and Pervasive Devices,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

³AIER Eye Hospital Group Co., Ltd., Changsha 410015, China
jiangxinlong@ict.ac.cn

Abstract

Medical AI benchmarks routinely pair a diagnostic image with the same-visit clinical record, delete the recorded diagnosis, and treat what remains as auxiliary context. The recipe is reasonable. It is not. On 5,904 ophthalmic visits across eight cohorts, a frozen text probe still recovers the disease label from the diagnosis-redacted record at 0.944 macro-AUROC, against 0.998 on the unredacted record; the most aggressive dictionary scrub closes less than one further percentage point. The diagnosis is already in the notes.

It is not in one place. A clinical record is written during the diagnostic process, and its fingerprints persist across channels: narrative text alone reaches 0.978, and a view that erases every value and exposes only which fields were populated still reaches 0.854. We introduce CRAFT (Channel-wise Redaction And Floor Test), a portable audit protocol that renders each record into redacted and channel-isolated views and reports each channel’s label recoverability; rankings reproduce under Qwen3.5-9B. Before clinical-text or image-text fusion gains are read as cross-modal reasoning, the text branch’s channel-level shortcut floors must be reported alongside them. <https://github.com/yhc98002-bit/The-Diagnosis-Is-Already-in-the-Notes>

1 Introduction

Medical AI is increasingly multimodal. A standard practice pairs a diagnostic image with the same-visit clinical record, strips the recorded diagnosis, and feeds the rest of the record to a text encoder as auxiliary context for image-based prediction (Zhang et al., 2023; Tu et al., 2024; Alsentzer

et al., 2019). The setting maps cleanly onto ophthalmology: a single visit that produces an OCT or fundus image also produces a record covering visual acuity, chief complaint, per-eye findings, orders, and the recorded diagnosis. Once the diagnosis fields are stripped, the remaining record is taken to be diagnostically neutral — clinical evidence to help the image branch, not to substitute for it.

It is not. On 5,904 multi-site ophthalmic visits across eight disease cohorts — five target diseases and three non-target clinical controls — a frozen-encoder text probe still recovers the disease label from the diagnosis-redacted record at 0.944 macro-AUROC, against 0.998 on the unredacted record. Throughout, that label is a retrospective cohort assignment made by the hospital network, not an independently adjudicated diagnosis (Section 3.4). The most aggressive surface scrub of the remaining text closes less than one further percentage point. The reason is structural. A clinical record is not a pre-diagnostic observation sheet — it is written during the diagnostic process itself: every choice the clinician makes — which tests to order, which drugs to prescribe, which fields to fill, which words to use — already reflects the diagnosis they are forming. Strip the diagnosis line, and that diagnosis has already written itself into everything else; even erasing the values of every measurement and leaving only which fields were filled still recovers a substantial fraction of the label. The diagnosis is already in the notes.

Two terms recur throughout, so we fix them here. We define a *shortcut* as label-predictive information a benchmark did not intend a model to use; a clinical channel becomes a shortcut only when a benchmark treats a same-visit, diagnosis-redacted record as neutral auxiliary context for evaluating image-based or image-text reasoning. We define

[†] Corresponding author.

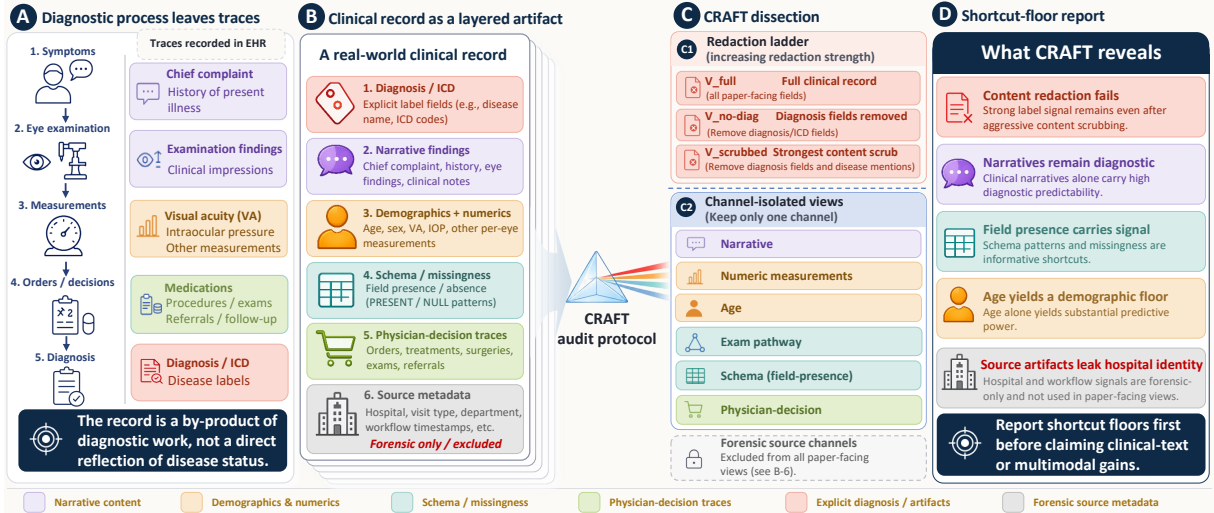


Figure 1: **Overview of CRAFT.** The diagnostic process leaves traces across the clinical record (A–B); CRAFT renders each record into redacted and channel-isolated views (C) and reports each channel’s shortcut floor (D). Forensic metadata is isolated from clinical views.

the *floor* (or *label floor*) of a view as the label recoverability achievable from that restricted view — a lower bound any claimed cross-modal gain must exceed. The channels audited here are not clinically meaningless; many are genuine by-products of care. CRAFT therefore does not ask whether the record is clinically useful, but what label floor the text branch already provides.

The scope follows from this. We audit the clinical-record branch and stop there: we train no image–text model, and so give no direct evidence for or against cross-modal reasoning in any particular system. The ordering is deliberate, since a fusion gain is interpretable only against a known floor. CRAFT is a prerequisite audit, not a verdict on multimodal reasoning.

Our contributions:

Protocol. We introduce CRAFT (Channel-wise Redaction And Floor Test), an audit protocol that decomposes a clinical record into a *redaction ladder* of progressively redacted views and a set of channel-isolated views, isolates cohort-construction artifacts in dedicated forensic views, and verifies every rendering under a fail-closed audit. CRAFT is portable: it needs only a record schema and view definitions, instantiable on any structured clinical-record corpus.

Finding. Diagnosis-field removal does not neutralize the record. On our 5,904-visit corpus, a frozen text probe still recovers the eight-way disease label at 0.944 macro-AUROC after the struc-

tured diagnosis fields are stripped, and 0.935 after a three-tier dictionary scrub — against 0.998 on the unredacted record.

Characterization. The label is not stored in one place. It leaks through narrative text, physician-decision traces, age alone, which fields were populated, and numeric measurements — *redundantly*, so single-point redaction does not remove it. Within the numeric channel, we additionally identify a missingness shortcut: zero-filling absent values makes absence itself diagnostic.

Guidance. Before interpreting any clinical-text or image-text fusion gain, report the text branch’s channel-level shortcut floors — at minimum text-only, schema-only, age-only, physician-decision, and source-metadata forensic — together with input-length statistics. Otherwise, a gain over an image-only baseline cannot be distinguished from the text remembering what the diagnosis used to be.

2 Related Work

Terminology. Our benchmark-relative use of *shortcut* and *floor* (Section 1) follows the dataset-artifact literature (Kaushik and Lipton, 2018; Tsuchiya, 2018). Cohort-construction metadata (source institution, export-file markers) is a separate category — never valid diagnostic input — and is isolated in dedicated forensic views.

Benchmark shortcuts, natural language understanding (NLU) artifacts, and probing. Shortcut learning was crystallized by Geirhos et al. (2020), with subsequent surveys mapping the resulting literature (Zhou et al., 2024). McCoy et al. (2019) showed that natural-language inference models rely on syntactic heuristics that fail on adversarial sets. A parallel line on dataset artifacts demonstrated that NLI labels can be predicted from the hypothesis alone (Gururangan et al., 2018; Poliak et al., 2018); Niven and Kao (2019) reported analogous effects in argument comprehension. Probing classifiers are a standard NLP analysis tool (Belinkov, 2022), and Bowman and Dahl (2021) argued that benchmark interpretation requires explicit baselines for what is recoverable from minimal cues (Lehman et al., 2023; Sainz et al., 2023). CRAFT applies this paradigm to routine clinical records: each channel-isolation view is a single-channel probe, and the resulting recoverability scores serve as shortcut floors.

Medical AI shortcuts and multimodal clinical pipelines. Medical-AI shortcut work has largely centered on imaging, including site markers in chest radiographs (Zech et al., 2018), COVID-19 dataset confounders (DeGrave et al., 2021), and latent race signals (Gichoya et al., 2022); clinical-text shortcuts have also been reported in the extraction of social determinants of health (SDOH) (Sakib et al., 2025). We ask the analogous question for the clinical-text branch increasingly paired with images in multimodal pipelines (Gulshan et al., 2016; De Fauw et al., 2018; Zhou et al., 2023; Zhang et al., 2023; Tu et al., 2024; Alsentzer et al., 2019).

De-identification versus label-neutralization. Clinical NLP has a long tradition of de-identifying free-text records to remove protected health information (Stubbs and Uzuner, 2015; U.S. Department of Health and Human Services, 2012; Lee et al., 2019; Johnson et al., 2016; Uzuner et al., 2011). De-identification removes patient identifiers; it is not designed to make a record label-neutral. A diagnosis-redacted note may be safe with respect to protected health information while still strongly predictive of the benchmark label. Our audit makes that gap measurable: removing diagnosis fields and applying additional dictionary scrubbing reduces label recoverability only marginally, because the diagnostic process is encoded in channels de-identification does not touch.

Cohort	Corpus n	Corpus %	Test n
Uveitis	2,354	39.9	225
AMD	446	7.6	47
DR	443	7.5	39
RP	404	6.8	42
RVO	383	6.5	32
Refractive (control)	680	11.5	68
Dry-eye (control)	661	11.2	63
Conjunctivitis (control)	533	9.0	56
Total	5,904	100	572

Table 1: Per-cohort visit counts in the corpus and in the test split.

3 Dataset

3.1 Source and scope

The corpus is a real-world ophthalmic image-clinical database compiled from a large multi-site ophthalmology hospital network operating across multiple Chinese provinces. It links diagnostic images to visit-level clinical records drawn directly from routine hospital information systems. Our object of study is the record itself: structured fields, free-text observations, and order entries produced during an ophthalmic consultation.

The records are routine clinical documentation, not a purpose-built benchmark (Zhang et al., 2022; Zhu et al., 2024). They were written for patient care, not as model inputs (Johnson et al., 2023).

3.2 Scale and cohorts

The corpus comprises 5,904 visit-level clinical records from 5,539 unique patients, organized into eight cohorts. Five cohorts correspond to target ophthalmic diseases — uveitis, age-related macular degeneration (AMD), diabetic retinopathy (DR), retinitis pigmentosa (RP), and retinal vein occlusion (RVO). The remaining three are non-target clinical control cohorts: refractive-error visits, dry-eye evaluations, and conjunctivitis presentations. They are not disease-free controls; they are walk-in ophthalmology populations that a deployed system must distinguish from the five target diseases. Table 1 reports the per-cohort record counts.

The corpus is class-imbalanced: uveitis accounts for roughly 40% of records, the other cohorts 6–12% each. We report macro-averaged and per-class AUROCs so smaller cohorts remain visible. The smallest cohorts have few test-split visits (RVO 32, DR 39, RP 42), so their per-class AUROCs are indicative rather than precise.

3.3 Structure of a clinical record

Each visit record combines structured fields (recorded diagnosis, ICD-10 code, demographics, per-eye measurements), free-text narrative (chief complaint, history, per-eye observations), and semi-structured orders (drugs, tests, treatments, surgeries, examinations) — the diagnostic conclusion, the evidence for it, and the actions taken in response, all in one record. Two of these structured fields are distinct and are treated separately throughout: the *recorded diagnosis* is the clinician’s free-text diagnostic statement as typed into the record, whereas the *ICD-10 code* is the coded classification assigned to that visit for administrative and billing purposes. Both are removed together by the redaction ladder (Section 4.3). This structure motivates the channel decomposition in Section 4.

3.4 Label construction and cohort assignment

The eight-class disease label is a *retrospective cohort assignment* made at data-acquisition time by the hospital network, not an independent diagnosis adjudication performed for this study. Labels were taken as-is from disease-stratified hospital exports, without re-adjudication or chart review. The structured diagnosis fields described above — the free-text recorded diagnosis and the ICD-10 code, with the related diagnostic fields — are exactly what the redaction ladder removes. They were attached to each record *after* the network had assigned the cohort, so they never entered label construction: removing them removes recorded diagnostic content, not the mechanism that produced the labels. Patients are nearly disjoint across cohorts (two pairwise overlaps in the entire corpus), consistent with disease-stratified hospital exports; export details are in Appendix E.

This retrospective cohort assignment has two direct consequences. First, the full-record AUROC is an expected ceiling: it measures recoverability from the complete documentation of the diagnostic episode that placed the patient in the cohort. Second, the primary object of study is the residual recoverability that persists *after* structured diagnosis fields are removed. It is that residual signal, not the ceiling, that the channel audit characterizes.

3.5 Patient-level split and privacy

We partition the corpus 8:1:1 at the patient level: all visits from a given patient go to the same partition

via a hashed patient key, yielding 4,734 / 598 / 572 training / validation / test visits. The split is fixed in advance and independent of the training seed, so every rendered view and every seed is evaluated on an identical partition. Records are keyed only by hashed identifiers; raw identifiers are absent from the analysis-ready data, and release details are in Appendix E.

4 Method

4.1 Probing task and setup

The task is an eight-way classification: given a rendered text view of a visit record, a classifier predicts the cohort.

Notation. Let x denote a clinical visit record and $y \in \{1, \dots, K\}$ its cohort label ($K = 8$). A **view** is a function v that maps a record to a rendered text string $v(x)$. CRAFT instantiates a **redaction ladder** — an ordered sequence of views of the same record in which each rung removes strictly more diagnostic content than the one below it — here $\{V_{\text{full}}, V_{\text{no-diag}}, V_{\text{scrubbed}}\}$ (Section 4.3), together with channel-isolation views $\{v_c\}_{c \in \mathcal{C}}$, each exposing a single information channel. The formal probe family and recoverability equation are in Appendix E; for a channel-isolation view v_c , its recoverability score is the **channel floor** for channel c under the probe.

Probes. h_{short} is chinese-roberta-wwm-ext (Cui et al., 2020), a compact Chinese encoder with a 512-token context window, building on the RoBERTa pre-training recipe (Liu et al., 2019) and the BERT architecture (Devlin et al., 2019). We use it for tractability across many views and seeds rather than alternatives such as ChineseBERT (Cui et al., 2021), MC-BERT (Zhang et al., 2020), or BGE-large-zh (Xiao et al., 2024). The long-context probe h_{long} is Qwen3.5-9B (Qwen Team, 2026). Both encoders are frozen throughout, so AUROC differences across views reflect rendered content, not model tuning.

4.2 CRAFT: a channel-wise clinical-record audit protocol

CRAFT (Channel-wise Redaction And Floor Test) is a clinical-record audit protocol with three steps: render each record into redacted and channel-isolated views, probe each view with a frozen-encoder classifier, and audit every rendering for intended content. Its three components are:

Record decomposition renders each record into (i) a *redaction ladder* that progressively removes diagnostic content, and (ii) a set of *channel-isolation views* that each expose a single information channel; **shortcut quantification** scores each view by the frozen-encoder probe, with the resulting AUROC measuring how readily the disease label can be recovered; and a **fail-closed rendering audit** checks every rendered view, for every visit, to confirm it emits its intended fields and no excluded content. Figure 1 gives an overview.

4.3 The redaction ladder

The redaction ladder applies progressively stronger content removal to the same record, producing three views: V_{full} renders the full clinical content of the record; $V_{\text{no-diag}}$ removes the *structured diagnosis fields* — recorded diagnosis text, ICD-10 code, and related diagnostic fields; and V_{scrubbed} further applies a stringent dictionary redaction over three tiers — (Tier 1) disease-entity mentions, (Tier 2) highly specific clinical findings, and (Tier 3) suggestive symptoms and historical context.

The scrubbing dictionaries are data-driven rather than a hand-built ontology such as CMeKG (Byambasuren et al., 2019): an automated PMI phrase-association pass plus a curated supplement yields 644 phrases across three tiers (Appendix A.4).

One property of this dictionary bounds how V_{scrubbed} should be read, so we state it here rather than in an appendix. V_{scrubbed} is *not* a deployable redaction pipeline: it is an analysis-time stress test on top of $V_{\text{no-diag}}$, and its phrase statistics were computed over the full 5,904-visit corpus, test split included — deliberately conservative, but it does let the dictionary see the evaluation data. Rebuilding it from the 4,734 training visits alone (629 versus 644 phrases; no validation or test patient enters the PMI pass) changes nothing: V_{scrubbed} moves 0.9349 to 0.9358 macro-AUROC and the scrubbed narrative 0.9649 to 0.9645 (Appendix F.3).

Throughout the ladder, every record is rendered under a *uniform schema*: each field appears in a fixed position, with an explicit null placeholder when absent. The ladder therefore measures removal of *content*, not changes in which fields are present; field presence is isolated below.

4.4 Channel-isolation views

In addition to the ladder, CRAFT renders channel-isolation views — each exposing a single channel,

so the disease signal carried by that channel alone can be measured. Full per-level definitions are in Appendix A. We report three families: the **content ladder** (V_{full} , $V_{\text{no-diag}}$, V_{scrubbed}) plus a label-shuffled control (text identical to V_{full} , training labels randomly permuted); **isolated clinical channels** covering narrative text, per-eye numerics, age, physician-decision fields, exam pathway, coarse per-eye status, and field presence with all values removed; and **forensic source channels** that expose only cohort-construction metadata (source institution, export-file markers) — *not valid diagnostic input*, removed from all non-forensic clinical views.

4.5 Training and evaluation protocol

For every view, the frozen encoder represents each rendered visit and a linear head is trained on the training split; optimizer, weighting, early-stopping, and extraction details are in Appendix E. We report test macro-AUROC and per-class AUROC as 3-seed mean $\pm$ std. These $\pm$ values are *seed stability* under a fixed split; they are not test-set confidence intervals and should not be read as such. Test-set uncertainty is reported separately as patient-level bootstrap 95% confidence intervals ($B=2000$ replicates, resampling unique patients), alongside macro-AUPRC for the class-imbalanced setting, in Table 2 and Appendix F.1. The chance reference for an eight-class macro-AUROC is 0.5; a label-shuffled control (training labels randomly permuted) serves as a near-chance pipeline sanity check.

Channel floors are not additive: channels share information, so their AUROCs do not sum to the full-record score. Each view asks whether the label is recoverable from a single channel, not how much that channel contributes to the full record. Context-window and truncation design, visual-acuity parsing, and per-view truncation rates are detailed in Appendices E and C.

5 Results

5.1 Content redaction does not remove the diagnosis

For diagnosis-redacted text to serve as auxiliary context in an image-clinical benchmark, deleting the diagnosis fields should sharply reduce the text branch’s standalone label recoverability. The redaction ladder shows the opposite: the full record reaches 0.998 macro-AUROC, while the label-shuffled control stays near chance at 0.476, confirm-

ing that the probe is reading genuine label structure.

Removing the structured diagnosis fields does not remove the diagnosis. Removing the recorded diagnosis text, ICD-10 code, and related structured diagnostic fields reduces macro-AUROC by only 0.054, leaving the probe at **0.944**. The explicit diagnostic entries exhibit high informational redundancy: once they are removed, the rest of the record remains highly discriminative.

As a secondary check, we apply an aggressive analysis-time dictionary scrub on top of diagnosis-field removal — masking disease-entity mentions, highly specific findings, and suggestive symptoms in the free text (V_{scrubbed} ; Section 4.3). This closes little of the remaining gap: macro-AUROC falls by only a further 0.009, to **0.935**.

The residual signal survives imbalance-sensitive metrics. The corpus is class-imbalanced and some test cohorts are small, so Table 2 reports macro-AUPRC and patient-level bootstrap 95% CIs for every view. Under AUPRC (chance 0.125) the shuffled control stays at 0.150 [0.146, 0.168] while $V_{\text{no-diag}}$ reaches 0.778 [0.739, 0.821] and the narrative 0.891 [0.856, 0.921]. Nor is the result carried by the majority class: on the three smallest test cohorts, per-class AUROC after diagnosis removal is 0.911 [0.851, 0.958] for DR, 0.936 [0.888, 0.974] for RP and 0.931 [0.884, 0.966] for RVO. Calibration of the same probe (Appendix F.2) characterizes the minimal linear head and is orthogonal to recoverability.

These full-record numbers use the legacy encoder under a 512-token window, which truncates the longer levels for many visits; the long-context check in Section 5.6 scores the same levels on essentially complete records and recovers the label at least as well. Appendix B reports per-class AUROCs at every level. The ladder therefore points to the central question for the channel audit: if diagnosis fields are not the main carrier of disease signal, which record channels are?

5.2 The disease label is distributed across record channels

We render each visit into single-channel views and probe them independently. Table 2 reports each channel as a shortcut floor, alongside the near-chance label-shuffled control (0.476) as a sanity check.

The narrative carries the diagnosis even after content scrubbing. Narrative text with field names stripped reaches 0.978 macro-AUROC, and the fully scrubbed narrative still reaches 0.965. The signal is not limited to disease names; dictionary scrubbing removes named phrases, but not the diagnostic language shaped by routine clinical reasoning.

Field completion patterns also carry disease signal. With every value removed and each field rendered only as PRESENT or NULL, the probe reaches 0.733, far above the 0.476 floor. Probing the same binary presence information three ways — RoBERTa text probe, long-context text probe, and direct logistic regression on the binary indicator vector — gives 0.733, 0.801, and 0.854 respectively (Figure 2; full setup in Section 5.6). The strongest result comes from direct binary features, indicating that the signal resides in the documentation pattern itself. Removing the five structured diagnosis-presence flags lowers the RoBERTa probe only marginally to 0.722 (Table 5) and the direct binary model to 0.834, so the schema signal is not driven by diagnosis-field presence itself. In that 39-field model the strongest presence coefficients are examination, order, laboratory-result, surgery and per-eye fields (Appendix F.4) — the documentation-completeness pattern of the clinical workflow, not residual diagnostic text.

Physician-decision traces are predictive on their own. Orders for drugs, laboratory tests, treatments, surgeries, and examinations reach 0.832 with no narrative, diagnosis, or raw results. A lab-result-only view reaches 0.583, and the older mixed orders-plus-results view reaches 0.835, so the signal is almost entirely in the decision fields. The selected drug, surgery, or test is itself diagnosis-shaped. Splitting the channel into its five sub-fields locates the signal more precisely: drug orders alone reach 0.779, ordered examination names 0.687, treatments 0.633, surgeries 0.558 and laboratory orders 0.549 (Appendix F.5). What is prescribed carries the diagnosis; what is merely requested carries much less of it. Additional clinical channels show the same pattern: the ordered-exam pathway reaches 0.765, and per-eye examination findings reduced to coarse normal/abnormal status reach 0.757 — both far above the shuffled-label floor.

Taken together, these views show that the disease label is not stored in any single place in the record. It is encoded redundantly — in the nar-

View	macro-AUROC		macro-AUPRC	
	3-seed mean	95% CI	point	95% CI
<i>Redaction ladder</i>				
Label-shuffled control	0.4760 ± 0.0815	[0.459, 0.493]	0.1505	[0.146, 0.168]
V_{full} — full record	0.9977 ± 0.0001	[0.996, 0.999]	0.9814	[0.971, 0.990]
$V_{\text{no-diag}}$ — diagnosis fields removed	0.9441 ± 0.0036	[0.930, 0.958]	0.7775	[0.739, 0.821]
V_{scrubbed} — strongest dictionary scrub	0.9349 ± 0.0038	[0.919, 0.949]	0.7297	[0.687, 0.772]
<i>Isolated clinical channels</i>				
Narrative only	0.9784 ± 0.0003	[0.970, 0.985]	0.8914	[0.856, 0.921]
Narrative, dictionary-scrubbed	0.9649 ± 0.0001	[0.954, 0.974]	0.8119	[0.772, 0.855]
Physician-decision only	0.8322 ± 0.0079	[0.808, 0.856]	0.4871	[0.451, 0.543]
Age + imputed VA/IOP numerics	0.7831 ± 0.0031	[0.756, 0.809]	0.3629	[0.342, 0.409]
Exam-pathway only	0.7646 ± 0.0009	[0.741, 0.790]	0.3621	[0.328, 0.410]
Per-eye coarse status	0.7568 ± 0.0052	[0.732, 0.781]	0.3609	[0.334, 0.404]
Schema field-presence only	0.7325 ± 0.0087	[0.706, 0.758]	0.2949	[0.274, 0.337]
Age only	0.7130 ± 0.0010	[0.691, 0.736]	0.2708	[0.257, 0.302]

Table 2: Disease-label recoverability from the redaction ladder and isolated clinical-record channels, patient-level split. The $\pm$ values are 3-seed standard deviations and measure *seed stability* under the fixed split. The 95% CIs are patient-level bootstrap intervals ($B=2000$, resampling unique patients) computed on the seed-ensembled test probabilities; their point estimates agree with the 3-seed means to ≤ 0.001 macro-AUROC. Chance macro-AUPRC is 0.125 for eight classes. Full intervals, bootstrap standard errors and per-class CIs are in Appendix F.1; source-metadata forensic views are reported separately in Section 5.5.

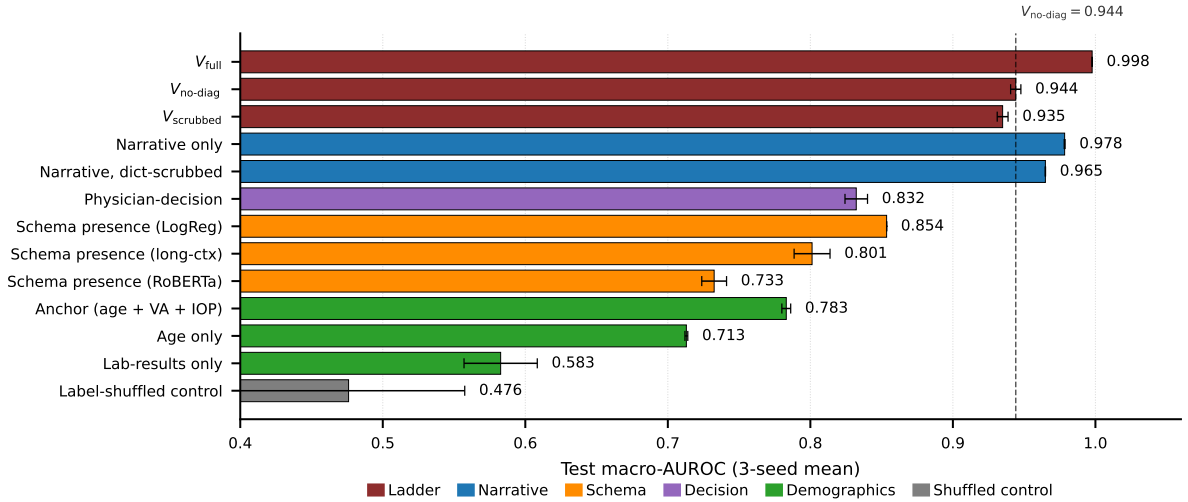


Figure 2: **Channel-floor map.** Macro-AUROC for the redaction ladder and channel-isolation views on a common scale. Dashed line marks $V_{\text{no-diag}}$ at 0.944: narrative alone (0.978) and the dictionary-scrubbed narrative (0.965) clear this floor without ever seeing the diagnosis. Bars are color-coded by channel family (legend in Figure 1); 3-seed mean with ± 1 standard-deviation error bars (seed stability under the fixed split, as in Table 2). Forensic source-metadata views are excluded (Section 5.5).

rative description, in which fields are populated, in the order patterns, in the examination pathway, in the per-eye findings — and several channels alone recover substantial signal. Single-point content redaction fails because removing one channel leaves the label recoverable from the others. The diagnosis is in all of them — its fingerprints, not its signature.

5.3 A demographic-prior floor

Patient age is a clinically meaningful covariate, not a cohort-construction artifact. Yet these cohorts have systematically different age distributions, so age alone becomes a strong disease predictor. Any benchmark built on such cohorts inherits this demographic-prior floor.

A classifier given only the patient’s age reaches 0.713 macro-AUROC. This accounts for most of

View	Legacy	Long-ctx
$V_{\text{no-diag}}$ — diagnosis fields removed	0.9441	0.9857
V_{scrubbed} — strongest dictionary scrub	0.9349	0.9723
Narrative only	0.9784	0.9888
Narrative, dictionary-scrubbed	0.9649	0.9842
Schema field-presence only	0.7325	0.8012
Physician-decision only	0.8322	0.8714

Table 3: Dual-probe comparison. Test macro-AUROC, 3-seed mean, matched batch size, patient-level split. Per-view tokenizer truncation rates are reported in Appendix C.

the conservative numeric anchor: age and the six per-eye VA/IOP measurements (median-imputed) reach 0.783, and removing age lowers it to 0.731. Age-only performance is uneven in exactly the expected way: it separates refractive-screening visits at 0.969 and AMD at 0.870, but reaches only 0.536 for retinitis pigmentosa.

Age-only performance should therefore be reported as a demographic-prior floor whenever clinical records or images are evaluated on cohorts with known age structure, and model performance interpreted relative to that floor.

5.4 A missingness shortcut in the numeric channel

The numeric measurements appear to be the cleanest channel: a small set of per-eye scalars, with no narrative, diagnosis, or order text. Even here, the diagnosis leaks — but through how the data is missing, not through what it says. In a zero-filled rendering, missing fields become explicit machine-readable values, and the numeric view reaches 0.796 ± 0.007 . Replacing zero-fill with specification-correct median imputation lowers the same view to 0.783 ± 0.003 , the value reported in Table 2; the drop is consistent across seeds and exceeds seed variation.

The model was reading both values and absences. Zero-filling made absent fields recognizable — and absence itself encodes a clinical decision. Which measurements the clinician chose to order, and which they did not, is conditioned on what disease they already suspected; the pattern of empty fields is the clinician’s diagnostic hypothesis, written into the data by what was not measured. The diagnosis is in what was not measured. Median imputation removes the surface cue, but the deeper signal persists: even a nominally “clean” numeric channel inherits the diagnosis-shaped ordering that produced it.

View	8-class	5-class
V_{full}	0.998	0.996
$V_{\text{no-diag}}$	0.944	0.904
V_{scrubbed}	0.935	0.881
Schema presence only	0.733	0.686
Anchor (age, sex, VA, IOP; no imputation)	0.766	0.729

Table 4: Cohort restriction. Macro-AUROC, 3-seed mean; 5-class drops the three non-target controls.

5.5 Forensic audit: cohort-construction artifacts were excluded

Not all recoverable signal is clinical. A clinical-record corpus also carries traces of how its cohorts were assembled and exported — not valid diagnostic input — which we quantify in forensic views to make the exclusion visible.

The clearest case is the visit-type metadata field, which is 100% null for four of the five disease cohorts and recorded as the same outpatient value for all three control cohorts, reflecting the upstream export source rather than any clinical property of the visit. Visit-type alone reaches 0.837, hospital identity alone 0.842, and the full source-metadata bundle 0.954 — all forensic views, isolated and excluded from non-forensic clinical views (Table 6, Appendix A.2). A fail-closed rendering audit confirms zero violations across all 342,432 view-visit renderings. These forensic results quantify cohort-construction artifacts that non-forensic clinical views remove, not clinical documentation signals that remain after redaction.

5.6 Dual-probe agreement and the prefix-artifact test

We re-score six representative views with the 9B-parameter long-context encoder, at a matched batch size and the same patient-level split. The decisive comparison is the truncation contrast. Under RoBERTa, $V_{\text{no-diag}}$ is truncated for 66.2% of visits and V_{scrubbed} for 57.6%; under the long-context encoder, the same levels are truncated for only 4.5% and 2.1%. The long-context encoder therefore scores these levels on essentially complete records — and its macro-AUROC is not lower but higher (0.986 and 0.972 versus 0.944 and 0.935; Table 3). An encoder that sees the near-complete record recovers the label at least as well; the full-record signal therefore extends beyond the leading prefix.

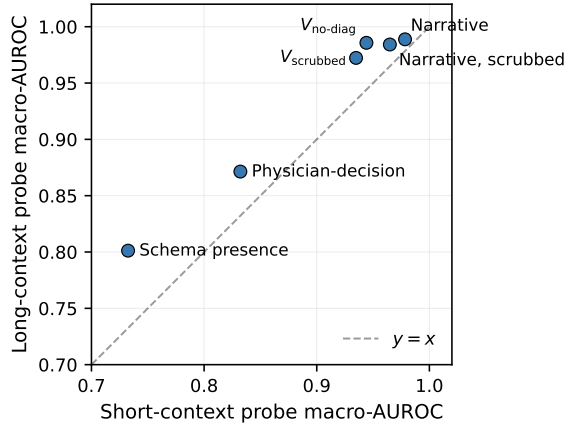


Figure 3: **Dual-probe agreement.** All six views re-scored under both probes (Table 3) lie on or above the identity line: the long-context probe recovers the label at least as well as the short-context probe despite heavy truncation (Table 7).

Schema feature set	n_{fields}	AUROC
All non-source-meta	44	0.733
No diag-presence flags	39	0.722
Common-core (all cohorts)	8	0.635
Label-shuffled control	—	0.476

Table 5: Schema restriction (8-class). Macro-AUROC, 3-seed mean.

All six re-scored views lie at or above the identity line (Figure 3), so the tested channel conclusions are stable across a short-context RoBERTa probe and a long-context 9B probe.

A non-textual tabular baseline for the field-presence channel. The field-presence result is the one most plausibly attributable to how presence is serialized into text, so we test it directly with a tabular baseline. We fit a logistic regression on the 44-dimensional binary field-presence vector — no text encoder, no tokenizer — using the same patient-level split. It reaches 0.854, above both the RoBERTa text probe (0.733) and the long-context text probe (0.801). Direct binary presence features expose the schema signal more cleanly than serialized text probes. Removing the fields that are constant across the training split — which include four of the five structured diagnosis-presence flags — leaves the result unchanged (0.854). The field-presence signal is therefore a property of the documentation pattern itself, not of any text encoder.

The dual-probe and tabular checks together show that Sections 5.1–5.2 reflect record structure, not probe-specific serialization.

5.7 Sensitivity to cohort and schema restriction

Restricting the corpus to the five disease cohorts alone (a five-class within-disease task), the ladder preserves its qualitative shape (Table 4): V_{full} 0.996, $V_{\text{no-diag}}$ 0.904, V_{scrubbed} 0.881; the field-presence view reaches 0.686 and the conservative numeric anchor 0.729. Thinning the schema from 44 to 39 to 8 common-core fields drops field-presence recoverability from 0.733 to 0.722 to 0.635 (Table 5) — near the age-only floor, indicating the documentation pattern carries diagnostic structure independent of any specific cohort-distinguishing field.

6 Discussion

Clinical text encodes the diagnosis not despite being genuine clinical evidence, but because of it. Narrative, numeric measurements, and examination findings are the clinician’s diagnostic reasoning made explicit; removing the diagnosis field leaves the others in place. Content redaction cannot neutralize such a record. CRAFT renders records into redacted and channel-isolated views and reports what each channel alone can recover: audit, not neutralization, as an explicit floor for downstream benchmarks.

This is not an argument for excluding clinical text from multimodal systems. The narrower point is that image–text gains must be read against the text branch’s floor before being attributed to cross-modal reasoning, and the mechanism is general: comparable channels arise wherever documentation and clinical workflow are conditioned on disease (Arjovsky et al., 2019; Sagawa et al., 2020; Recht et al., 2019).

A measured floor is only useful if it changes practice, which we take to mean three things. *Report*: five channel floors — text-only, schema-only, age-only, physician-decision and source-metadata forensic — belong beside any image–text gain, so the gain reads as the margin above them. *Baseline*: a text-only and channel-isolated baseline belongs in the evaluation table next to the image-only baseline; a fusion number without both is uninterpretable. *Constrain*: when a channel’s stand-alone floor approaches the multimodal score, exclude it from the benchmark input or constrain it through channel-level gating rather than crediting the result to cross-modal reasoning. The diagnosis was never only in the diagnosis line. It was in the notes.

Limitations

Single health system, and no cross-site generalization test. The corpus is multi-site but drawn from one hospital network, electronic-record system, and documentation template family. The exact AUROC values are therefore specific to this health system and should not be read as estimates for other hospitals, record systems, documentation templates, or specialties; the transferable contribution is the audit protocol and the failure mode it exposes. We also cannot settle the residual-confounding question empirically here. Removing explicit source-metadata fields does not remove indirect site, template, or export-pipeline style differences, and a clean leave-one-site-out or hospital-balanced evaluation is structurally infeasible on this corpus because site and cohort are strongly entangled in the disease-stratified export (Appendix F.6). Transparent forensic isolation is the control we can run. Cross-system replication therefore remains necessary to disentangle documentation-style effects from clinical-content effects, and we release the protocol so that it can be run.

Probing setup. A frozen encoder with a linear head measures how readily the disease label can be extracted under a controlled, fixed probe; it does not measure every form of information a record could in principle contain. This controlled setup makes the views comparable, but the reported numbers should be read as extractability under that probe, not as an absolute upper bound. For the same reason, the calibration of the probe (Appendix F.2) characterizes the minimal linear head rather than the record, and is not part of the floor argument.

Label and cohort construction. The eight-class labels are cohort assignments from the hospital network’s data-extraction process (Section 3.4), not independent prospective disease adjudications. We therefore state all results as recoverability of *retrospective cohort labels*, not of independently adjudicated diagnostic truth: cohort assignment, documentation practice, and export mechanics may each contribute to the measured signal, and the audit is designed to expose exactly that. The labels should not be read as an external diagnostic gold standard.

Redaction dictionary. The V_{scrubbed} dictionary scrub removes surface phrases identified by phrase-association analysis; it is not semantic rewriting and cannot remove compositional clinical evidence

expressed without those phrases. The dictionary is used only as an analysis-time stress test (Section 4.3), not as a deployable de-identification pipeline, and its phrase statistics were computed over the full analysis corpus; reconstructing it from the training split alone leaves the conclusions essentially unchanged (Section 4.3, Appendix F.3).

Text-branch-first scope. This paper intentionally audits the clinical-record branch before image-text fusion. It does not evaluate image-text fusion, laterality grounding, or whether multimodal systems retain gains once these floors are accounted for. We therefore make no claim, positive or negative, about cross-modal reasoning in any deployed or benchmark system: CRAFT is a prerequisite auditing framework that establishes the floor such a claim would have to clear, and a high floor is evidence that a multimodal gain is *uninterpretable* without it, not evidence that the gain is spurious. Measuring the residual gain above these floors is the direct next step, and we outline it in Appendix G.

Ethics Statement

The clinical records used in this study were collected during routine ophthalmic care within a multi-site hospital network. They were accessed for research under institutional authorization and a data-use agreement governing their handling, de-identification, and release.

Every record is identified only by a hash of the patient identifier; raw patient identifiers, hospital-internal record numbers, and raw file-path metadata are absent from the analysis-ready records and are not read by any audit or modeling script.

A controlled set of source-metadata fields — such as hospital identity and visit-type markers — is used only in the forensic probes of Section 5.5 and is excluded from every non-forensic clinical view. A fail-closed rendering audit verifies, across all 342,432 view-visit renderings, that no such source-metadata field is emitted in any non-forensic clinical view. We do not release the raw corpus; we release de-identified illustrative examples, the field schema, the rendering specifications, and the analysis code. The forensic views are reported only to quantify artifacts that the non-forensic clinical views exclude, and contain no information about individual patients.

Acknowledgments

We thank the three anonymous reviewers and the area chair for detailed and constructive comments, and in particular for the requests that led to the patient-level bootstrap confidence intervals, the train-only redaction dictionary, and the calibration, schema, and physician-decision analyses reported in Appendix F. We thank the clinical and data teams of the AIER Eye Hospital Group for assembling and governing the record corpus under the data-use agreement that made this audit possible.

References

- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. [Publicly available clinical BERT embeddings](#). *CoRR*, abs/1904.03323.
- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. [Invariant risk minimization](#). *arXiv preprint arXiv:1907.02893*.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Comput. Linguistics*, 48(1):207–219.
- Samuel R. Bowman and George E. Dahl. 2021. [What will it take to fix benchmarking in natural language understanding?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4843–4855. Association for Computational Linguistics.
- Odmaa Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. 2019. [Preliminary study on the construction of Chinese medical knowledge graph](#). *Journal of Chinese Information Processing*, 33(10):1–7.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. [Revisiting pre-trained models for Chinese natural language processing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, page 657–668. Association for Computational Linguistics.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. [Pre-training with whole word masking for Chinese BERT](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, and 15 others. 2018. [Clinically applicable deep learning for diagnosis and referral in retinal disease](#). *Nature Medicine*, 24(9):1342–1350.
- Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. [AI for radiographic COVID-19 detection selects shortcuts over signal](#). *Nat. Mach. Intell.*, 3(7):610–619.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nat. Mach. Intell.*, 2(11):665–673.
- Judy Wawira Gichoya, Imon Banerjee, Ananth R. Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, and 4 others. 2022. [AI recognition of patient race in medical imaging: a modelling study](#). *The Lancet Digital Health*, 4(6):e406–e414.
- Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. 2016. [Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs](#). *JAMA*, 316(22):2402.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. [Annotation artifacts in natural language inference data](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific Data*, 10(1).

- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3(1).
- Divyansh Kaushik and Zachary C. Lipton. 2018. [How much reading does reading comprehension require? a critical investigation of popular benchmarks](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 5010–5015. Association for Computational Linguistics.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. [Do we still need clinical language models?](#) In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4658–4664. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 180–191.
- Qwen Team. 2026. [Qwen3.5: Towards native multimodal agents](#). Qwen Team blog, <https://qwen.ai/blog?id=qwen3.5>.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. [Do ImageNet classifiers generalize to ImageNet?](#) In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5389–5400.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. [Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization](#). In *International Conference on Learning Representations (ICLR)*.
- Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. [NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.
- Fardin Ahsan Sakib, Ziwei Zhu, Karen Trister Grace, Meliha Yetisgen, and Ozlem Uzuner. 2025. [Spurious correlations and beyond: Understanding and mitigating shortcut learning in SDOH extraction with large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL) – Short Papers*. 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.
- Junyi Sun. 2012. [jieba: Chinese text segmentation](#). Open-source software, <https://github.com/fxsjy/jieba>.
- Masatoshi Tsuchiya. 2018. [Performance impact caused by hidden bias of training data for recognizing textual entailment](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, and 29 others. 2024. Towards generalist biomedical AI. *NEJM AI*, 1(3).
- U.S. Department of Health and Human Services. 2012. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) Privacy Rule. <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/>. Office for Civil Rights, accessed 2026-05-20.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. [2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text](#). *Journal of the American Medical Informatics Association*, 18(5):552–556.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-Pack](#):

Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 641–649. ACM.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. 2018. [Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study](#). *PLOS Medicine*, 15(11):e1002683.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, and 4 others. 2022. [CBLUE: A Chinese biomedical language understanding evaluation benchmark](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) – Long Papers*, pages 7888–7915. Association for Computational Linguistics.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. [Conceptualized representation learning for Chinese biomedical text mining](#). *arXiv preprint arXiv:2008.10813*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, and 2 others. 2023. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robert R. Struyven, Timing Liu, Moucheng Xu, Matteo G. Lozano, Peter Woodward-Court, Yuka Kihara, Naomi Allen, John E. J. Gallacher, Thomas Littlejohns, Tariq Aslam, Paul Bishop, Graeme Black, Panagiotis Sergouniotis, Denize Atan, and 75 others. 2023. [A foundation model for generalizable disease detection from retinal images](#). *Nature*, 622(7981):156–163.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. [Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614. Association for Computational Linguistics.

Wei Zhu, Xiaoling Wang, Mosha Chen, and Buzhou Tang. 2024. [Overview of the PromptCBLUE shared task in CHIP2023](#). In *Health Information Processing – CHIP 2023 Revised Selected Papers, Communications in Computer and Information Science*.

A Full level definitions and Level IDs

Each row in Tables 2 and 3 corresponds to a rendered view of the clinical record. The main-text name and the internal Level ID below are given side by side; the Level ID is provided for cross-reference with the released analysis code.

The internal Level IDs retain the naming of the released codebase, in which the redaction views developed through an l0/l1/. . ./l4 iteration. For the mapping to the main-text symbols: V_{full} corresponds to the l0 prefix, $V_{\text{no-diag}}$ to the l1 prefix, and V_{scrubbed} to the l4/L4 prefix; the three dictionary tiers (Tier 1/2/3) correspond to the l2/l3/l4 dictionary stages. The L4_common_core family of Level IDs denotes the numeric-channel views built on top of that scrub stage.

A.1 Levels reported in Tables 2 and 3

The conservative numeric anchor reported in Table 2 and discussed in Sections 5.3–5.4 (age and the six per-eye VA/IOP measurements, median-imputed, at 0.7831, and the same anchor without age at 0.7307) corresponds to the Level IDs `numeric_value_imputed_va_clean` and `numeric_value_imputed_no_age_va_clean`; their per-class values appear in Appendix B.

A.2 Forensic views

Forensic views expose only cohort-construction metadata (hospital identity, visit type, source-export markers); they are *not valid diagnostic input* and are excluded from every non-forensic clinical view in the main paper. They are reported here solely to quantify the cohort-construction artifacts that the protocol’s forensic isolation strips out (Section 5.5). The macro-AUROC for the three forensic views appears in Table 6 (companion to the truncation diagnostic in Table 7).

A.3 Supplementary channel-split levels

The physician-decision channel was constructed by splitting an earlier combined orders-plus-lab-result view into a pure physician-decision channel (`orders_decision_only`, Table 2’s “Physician-decision only”) and a complementary lab-result-evidence channel (`lab_results_only`, macro-AUROC 0.5827 ± 0.0257 , near-chance on its own). The unsplit legacy view (`orders_only`, macro-AUROC 0.8351 ± 0.0036) is retained only as a sensitivity check and is not used as a primary reported result.

Forensic view	macro-AUROC
visit_type only	0.8373 $\pm$ 0.0040
hospital identity only	0.8416 $\pm$ 0.0044
full source-metadata bundle	0.9540 $\pm$ 0.0004

Table 6: Forensic views — *not valid diagnostic input*; excluded from every non-forensic clinical level. Test macro-AUROC, 3-seed mean $\pm$ std. (Companion definitions in Table 10.)

A.4 Scrubbing-dictionary construction

The V_{scrubbed} scrubbing dictionaries were constructed by an automated phrase-association pass over the 5,904-visit analysis corpus, supplemented with a small hand-curated set. Visits were tokenized with jieba (Sun, 2012) into 2- to 12-character candidate phrases (covering the typical length of Chinese ophthalmic terminology and short descriptive findings); candidates appearing in fewer than 10 visits were discarded to filter out idiosyncratic phrasings and transcription typos, leaving 1,381 eligible phrases. For each eligible phrase we computed the maximum pointwise mutual information (PMI) across the eight cohorts under add-one smoothing, and retained phrases with maximum PMI at or above 1.0. This threshold was selected by sweeping a range of cutoffs and inspecting the resulting redaction strength: below 1.0 the retained phrases were near-uniformly distributed across cohorts and contributed little disease-specific masking, so 1.0 is the point past which additional phrases add scrubbing coverage without diluting the dictionaries with non-specific terms. Retained phrases were bucketed into the Tier 1 (disease-entity), Tier 2 (specific-finding), or Tier 3 (suggestive) dictionary by phrase length and field of occurrence. After deduplication and merging with the curated supplement, the three dictionaries together contain 644 phrases, broken down by tier in Table 8 below.

B Per-class AUROC matrix

For completeness, per-class one-vs-rest AUROCs for every view in Tables 2 and 3 are produced by the released analysis code from the per-seed result files. The full views $\times$ 8 matrix is large; we summarize the salient patterns rather than reprint every cell.

View	Short-ctx %	Long-ctx %
$V_{\text{no-diag}}$ — diag. removed	66.24	4.49
V_{scrubbed} — strongest scrub	57.57	2.13
Narrative only	0.02	0.02
Narrative, dictionary-scrubbed	0.02	0.02
Schema field-presence only	0.00	0.00
Physician-decision only	6.76	3.05

Table 7: Per-view fraction of visits whose pre-truncation token count exceeds the 512-token rendering budget under each encoder’s tokenizer. The two full-record views ($V_{\text{no-diag}}$ and V_{scrubbed}) are heavily truncated under the short-context probe but only marginally under the long-context probe; the channel-isolation views are short and not materially truncated under either probe.

Salient patterns. Across all paper-facing views, the three non-target control cohorts (refractive-error, dry-eye, conjunctivitis) are the most predictable from documentation-pattern channels (schema field-presence, exam pathway). The smallest disease cohorts (RVO with 32 test visits, DR with 39, RP with 42) show wider per-seed variability in their one-vs-rest AUROC than the macro-averaged figures in Table 2; per-class values on these cohorts should be read as indicative rather than precise. The label-shuffled control remains near 0.5 in every per-class slot, confirming that the patient-level split and class-weighted training procedure do not introduce a systematic baseline above chance.

Per-class numbers under the long-context probe (Table 3 views) reproduce the same ordering as the RoBERTa probe on every cohort tested.

The full per-class matrices are released alongside the analysis code for reproducibility.

C Per-view tokenizer truncation rates

The two probe encoders use different tokenizers and context windows. We report the fraction of visits whose pre-truncation token count exceeds the 512-token rendering budget, separately for each tokenizer. The short-context probe (chinese-roberta-wwm-ext) uses the Chinese RoBERTa WordPiece tokenizer; the long-context probe (Qwen3.5-9B) uses the Qwen BPE tokenizer.

The short-context probe therefore measures the diagnosis-fields-removed and strongest-scrub views from the leading prefix of each record; the long-context probe reads essentially complete records on the same views. Dual-probe agreement under this truncation contrast (Table 3) is what establishes that the full-record recoverability

Dictionary tier	Target phrases	Auto	Curated	Total
Tier 1 — disease-entity	disease names and entities	200	71	267
Tier 2 — specific-finding	disease-specific examination findings	71	113	177
Tier 3 — suggestive	suggestive symptoms and history	200	0	200
Total		471	184	644

Table 8: Data-driven redaction dictionaries used to build V_{scrubbed} . *Auto* phrases come from a PMI-thresholded phrase-association pass; *Curated* entries are a hand-added supplement; *Total* is the deduplicated phrase count loaded at scrub time.

extends beyond the leading prefix.

D Reference tables

The following tables provide the mapping from main-text view names to the internal Level IDs used in the released analysis code. Table 9 covers the clinical (non-forensic) views; Table 10 covers the forensic views.

E Method and Dataset Details

Cohort export mechanics. For the uveitis cohort, all records were drawn from a dedicated uveitis archive directory in the hospital network. For each of the other seven cohorts, the hospital network provided a per-disease patient-key whitelist (one spreadsheet per cohort) and the corresponding records were extracted from a shared imaging-and-clinical pool. The structured diagnosis fields removed by the redaction ladder — the free-text recorded diagnosis, the ICD-10 code, and the related diagnostic fields — were joined onto each record only *after* the cohort had been assigned, and were never used to assign it.

Privacy and release. Records are keyed by hashed patient identifiers, and raw identifiers are absent from the analysis-ready data. We do not release the corpus; we release de-identified examples, the field schema, rendering specifications, and analysis code so the protocol can be reproduced on other corpora.

Probe family and recoverability score. Each view is scored by a probe from a **probe family** $\mathcal{H} = \{h_{\text{short}}, h_{\text{long}}\}$, where h_{short} is a compact short-context encoder and h_{long} is a long-context encoder. A linear classification head g_h is trained on top of each encoder. The **recoverability score** of view v under probe h is

$$S^h(v) = \text{MacroAUROC}(g_h \circ h \circ v, \mathcal{D}_{\text{test}}),$$

the macro-averaged one-vs-rest AUROC on the patient-disjoint test split. For a channel-isolation

view v_c , $S^h(v_c)$ is the **channel floor** for channel c under probe h .

Training details. For every view, the frozen encoder produces a representation of each rendered visit, and a linear head is trained on the training split with a class-weighted cross-entropy loss (weights inverse to training-set class frequency). We use the AdamW optimizer, train for up to ten epochs with early stopping on validation macro-F1, and run every configuration with three random seeds; the seed affects only head initialization and minibatch order, not the data split. Test features are extracted once, after training completes, so the test split is touched a single time at final reporting.

Context window and truncation. The two probes differ sharply in input capacity, and this is central to the design. Under the RoBERTa encoder’s 512-token window, the full-record views are long enough to be truncated to their leading content for a large fraction of visits — on $V_{\text{no-diag}}$ and V_{scrubbed} , the majority of records exceed the window (Section 5.6). A probe restricted to that window therefore measures only what fits in the leading prefix. The 9B-parameter long-context encoder represents the same levels with only a few percent of visits truncated, so dual-probe agreement establishes that the full-record conclusion holds in both context-window regimes. The channel-isolation views are short and not materially truncated under either encoder; per-view truncation rates are reported in Section 5.6 and Appendix C.

Visual-acuity parsing. Ophthalmic records express very low vision with categorical notations — counting-fingers or hand-motion at a stated distance, light-perception, no-light-perception — rather than decimal acuity. We parse these as a categorical low-vision channel and retain only genuine decimal values in the numeric acuity fields, so a categorical low-vision finding is not

Name in Tables 2 and 3	Level ID	Definition
V_{full} — full record	10_clinical_no_source_meta	Full clinical content with source-metadata fields dropped; the ceiling of text-based predictability under the uniform-schema rendering.
$V_{\text{no-diag}}$ — diagnosis fields removed	11_clinical_no_source_meta	The full-record view after stripping the five structured diagnosis fields (diagnosis.raw, diagnosis.text, diagnosis.icd10, diagnosis.other, diagnosis.laterality).
V_{scrubbed} — strongest dictionary scrub	14_rich_clinical_no_source_meta	The $V_{\text{no-diag}}$ view further passed through the Tier 1 (disease-entity), Tier 2 (specific-finding), and Tier 3 (suggestive) dictionary scrubs over the narrative and retained fields.
Narrative only	narrative_only_no_fieldnames	Free-text narrative blobs (chief complaint, past history, per-eye observations) rendered as raw text with all field-name labels stripped.
Narrative, dictionary-scrubbed	narrative_only_L4	The narrative-only view after the full Tier 1 + Tier 2 + Tier 3 dictionary scrub.
Schema field-presence only	schema_only_all_no_source_meta	All 44 non-source-meta fields rendered as binary field: PRESENT / field: NULL; field <i>values</i> are stripped, presence indicators retained for all non-source fields including structured-diagnosis presence flags.
Age + imputed VA/IOP numerics	numeric_value_imputed_va_clean	Seven positional numerics — demographics.age and the six per-eye numerics (left/right va, corrected_va, iop) — with missing values imputed by train-split medians; visual acuity parsed with the categorical low-vision channel separated out. demographics.sex is not included in this view.
Physician-decision only	orders_decision_only	Five physician-decision fields: exam_orders.exam_name, drug_orders, lab_orders, treatment_orders, surgery_orders. No narrative, no diagnosis fields, no schema scaffold, no raw lab values.
Exam-pathway only	exam_pathway_no_counts	exam_orders.exam_name + exam_orders.has_result + modalities (three fields; integer file/device counts dropped).
Per-eye coarse status	eye_status_only_outpatient	Twenty-two per-eye anatomical fields rendered as coarse status (normal/abnormal/unknown) via the negation-aware classifier; restricted to outpatient visits after the split to avoid an inpatient-schema gap.
Age only	age_only	A single feature: demographics.age. The demographic-prior floor.
Label-shuffled control	ctrl1a	Text identical to the full record but with training labels shuffled within each split deterministically per seed; gives the near-chance level reachable under the pipeline.

Table 9: Mapping from main-text view names to internal Level IDs in the released code, with field-level definitions.

Name in Table 6	Level ID	Definition
Visit-type field alone	visit_type_only_sanitized	The sanitized demographics.visit_type field alone. Forensic; reported only to quantify the cohort-construction artifact.
Hospital identity alone	hospital_name_only	The hospital_name field alone. Forensic.
Full source-metadata bundle	source_meta_only	Six source-metadata fields together: hospital_name, hosp_id, encounter_kind, sanitized demographics.visit_type, visit_year, visit_no_kind. Forensic.

Table 10: Forensic view Level ID definitions.

misread as a numeric acuity value.

F Supplementary analyses

This appendix reports the additional analyses requested during review. All were produced by the released code from the same fixed patient-level split and the same three seeds (42/137/2024) as the main results; per-analysis reproduction commands, machine-readable JSON, and per-stage logs are released alongside the paper.

F.1 Patient-level bootstrap confidence intervals and macro-AUPRC

Test-set uncertainty is estimated by resampling *unique patients* with replacement ($B = 2000$ replicates, RNG seed 20260709); every test row belonging to a sampled patient is included, so the interval respects the patient-level structure of the split. Intervals are computed on the seed-ensembled test probabilities (arithmetic mean across the three seeds per record), whose macro-AUROC agrees with the 3-seed means reported in the main text to within 0.001. Table 11 gives the full intervals and bootstrap standard errors; no replicate dropped a class in any view.

Small cohorts. The three smallest test cohorts (DR 39, RP 42, RVO 32 visits) are the ones for which a macro-average could hide instability. Their per-class one-vs-rest AUROCs after diagnosis removal are 0.9110 [0.8511, 0.9585] for DR, 0.9362 [0.8875, 0.9738] for RP and 0.9309 [0.8835, 0.9664] for RVO; on the narrative-only view they are 0.9725 [0.9482, 0.9901], 0.9780 [0.9565, 0.9940] and 0.9691 [0.9372, 0.9918]. Every interval excludes chance by a wide margin, so the residual recoverability reported in Section 5.1 is not produced by the large uveitis cohort alone.

F.2 Calibration

Calibration is computed on the same seed-ensembled probabilities: top-label expected calibration error (ECE) over 15 equal-width confidence bins, classwise ECE as the macro mean of eight one-vs-rest reliability curves, and the multiclass Brier score (Table 12). Temperature scaling was not applied, since the probe is deliberately minimal and no validation probabilities were retained. These numbers characterize the frozen encoder plus linear head, not the record: a view can be highly recoverable and poorly calibrated at the same time, which is why the floor argument is stated in terms of AUROC and AUPRC rather than probability quality. Figure 4 shows reliability diagrams for four representative views.

F.3 Train-only redaction dictionary

The V_{scrubbed} dictionary of Appendix A.4 computes its phrase statistics over the full analysis corpus, which is a conservative stress test but does let the dictionary see the evaluation split. We therefore rebuilt it using only the 4,734 training visits: zero validation or test patient keys enter the PMI pass, yielding 629 merged phrases against 644 for the full-corpus dictionary (disease-entity 256 vs. 267; specific-finding 173 vs. 177; suggestive 200 vs. 200). Curated supplements were copied unchanged. Re-scoring the two affected views with the train-only dictionary moves V_{scrubbed} from 0.9349 to 0.9358 macro-AUROC ($\Delta = +0.0009$; AUPRC 0.7297 to 0.7307) and the scrubbed narrative from 0.9649 to 0.9645 ($\Delta = -0.0004$; AUPRC 0.8119 to 0.8087). Both deltas are an order of magnitude below the 0.054 drop the diagnosis-field removal produces, so the full-corpus dictionary is a conservative analysis-time choice rather than a source of leakage.

View	macro-AUROC (SE)	macro-AUPRC (SE)
V_{full}	0.9977 (0.0006)	0.9814 (0.0047)
$V_{\text{no-diag}}$	0.9441 (0.0072)	0.7775 (0.0213)
V_{scrubbed}	0.9349 (0.0074)	0.7297 (0.0219)
Narrative only	0.9784 (0.0038)	0.8914 (0.0166)
Narrative, scrubbed	0.9649 (0.0050)	0.8119 (0.0213)
Physician-decision	0.8325 (0.0120)	0.4871 (0.0230)
Age + VA/IOP	0.7831 (0.0135)	0.3629 (0.0170)
Exam pathway	0.7648 (0.0125)	0.3621 (0.0211)
Per-eye coarse status	0.7568 (0.0128)	0.3609 (0.0180)
Schema field-presence	0.7325 (0.0133)	0.2949 (0.0165)
Age only	0.7130 (0.0115)	0.2708 (0.0113)
Label-shuffled control	0.4760 (0.0086)	0.1505 (0.0057)

Table 11: Patient-level bootstrap point estimates and standard errors, $B = 2000$. The corresponding 95% intervals are in Table 2. Chance macro-AUPRC is 0.125.

View	top-label ECE	classwise ECE	Brier
V_{full}	0.3588	0.0876	0.2462
$V_{\text{no-diag}}$	0.3010	0.0867	0.4679
V_{scrubbed}	0.2910	0.0787	0.4790
Narrative only	0.1608	0.0535	0.2689
Narrative, scrubbed	0.1410	0.0543	0.3273
Physician-decision	0.1851	0.0779	0.6612
Exam pathway	0.1155	0.0705	0.7326
Per-eye coarse status	0.2095	0.0795	0.7834
Schema field-presence	0.1340	0.0701	0.8209
Age + VA/IOP	0.0742	0.0752	0.7393
Age only	0.0603	0.0762	0.7861
Label-shuffled control	0.0607	0.0790	0.8795

Table 12: Calibration of the frozen probe. $N = 572$ test visits for every view except per-eye coarse status ($N = 554$, outpatient-only by construction).

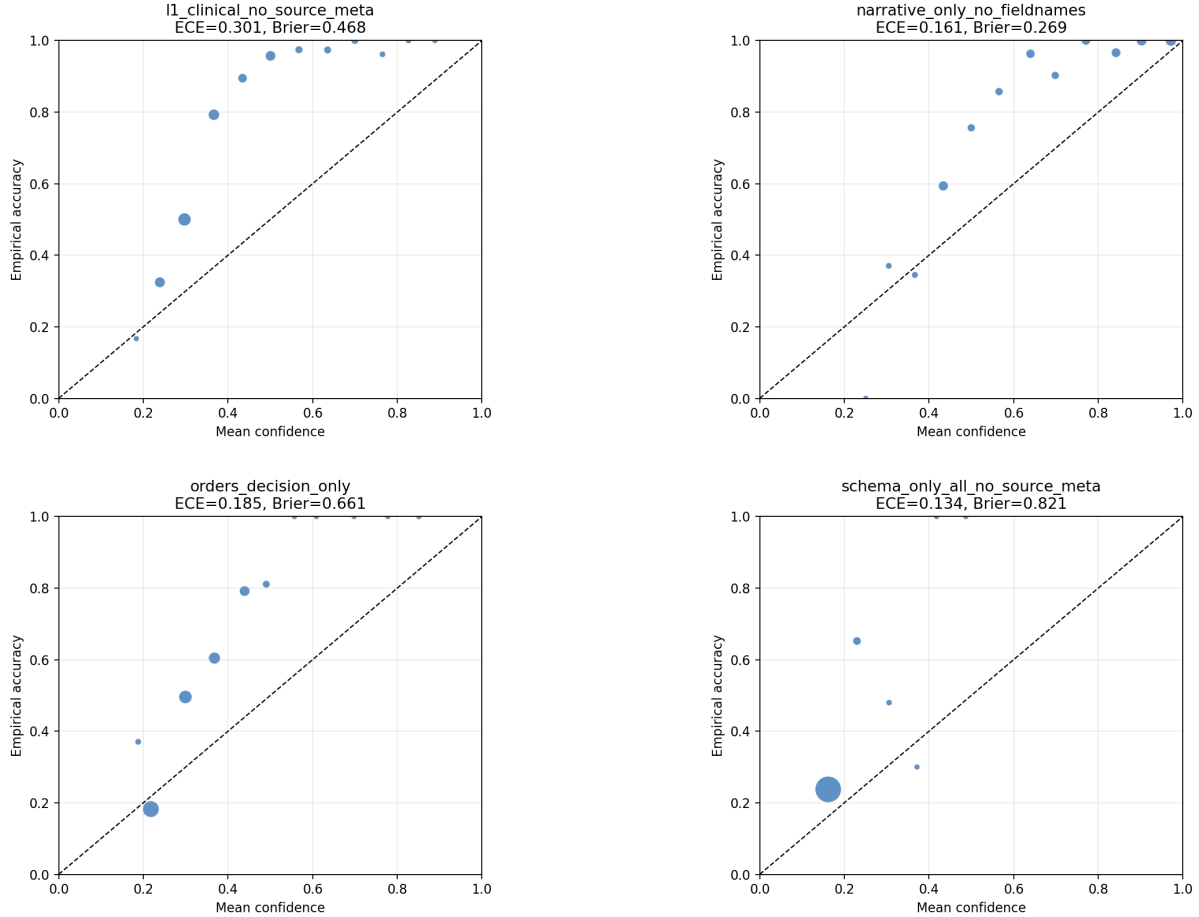


Figure 4: **Reliability diagrams.** Top: $V_{\text{no-diag}}$ (left) and narrative only (right). Bottom: physician-decision only (left) and schema field-presence only (right). Fifteen equal-width confidence bins on the seed-ensembled test probabilities; the dashed diagonal is perfect calibration and marker area is proportional to bin population.

F.4 What the field-presence channel reads

To interpret the field-presence floor we fit a multinomial logistic regression with L2 regularization and balanced class weights directly on the bi-

nary PRESENT/NULL matrix (Table 13), and inspect its coefficients after the five structured diagnosis-presence flags (`diagnosis.raw`, `.text`, `.icd10`, `.other`, `.laterality`) are removed.

Probe / field set	n_{fields}	AUROC	macro-F1
LogReg, all schema fields	44	0.8535	0.460
LogReg, diag. flags removed	39	0.8337	0.412
LogReg, non-constant fields	37	0.8540	0.460
RoBERTa, diag. flags removed	39	0.7216	—

Table 13: Field-presence recoverability under direct binary features and under the text probe, with and without the structured diagnosis-presence flags.

Sub-field	macro-AUROC	macro-AUPRC
Drug orders	0.7790 ± 0.0029	0.3688 ± 0.0032
Examination name	0.6869 ± 0.0080	0.2798 ± 0.0020
Treatment orders	0.6330 ± 0.0241	0.2651 ± 0.0365
Surgery orders	0.5580 ± 0.0394	0.2159 ± 0.0633
Laboratory orders	0.5493 ± 0.0528	0.2148 ± 0.0017
Combined channel	0.8322	0.4871

Table 14: Physician-decision sub-fields, 3-seed mean $\pm$ std. The bottom row repeats the combined orders_decision_only channel exactly as reported in Table 2: its macro-AUROC is the 3-seed mean and its macro-AUPRC is the seed-ensembled point estimate, so neither carries a per-seed spread.

Cohort	Top three positive presence fields	Strongest negative field
Uveitis	drug_orders (+0.74), right_eye.vitreous (+0.63), abnormal_lab_results (+0.63)	exam_orders.exam_name (−3.21)
AMD	right_eye.conjunctiva (+1.51), surgery_orders (+1.12), exam_orders.exam_name (+0.92)	right_eye.cornea (−2.65)
DR	abnormal_lab_results (+0.98), left_eye.cornea (+0.94), right_eye.va (+0.82)	left_eye.corrected_va (−1.30)
RP	chief_complaint (+0.77), right_eye.iop (+0.66), right_eye.cornea (+0.64)	right_eye.other (−0.98)
RVO	surgery_orders (+1.35), abnormal_lab_results (+0.92), lab_orders (+0.78)	right_eye.conjunctiva (−1.16)
Refractive	right_eye.cornea (+2.21), exam_orders.exam_name (+0.95), right_eye.conjunctiva (+0.95)	treatment_orders (−2.06)
Dry-eye	right_eye.other (+1.87), left_eye.other (+1.76), right_eye.iop (+0.92)	abnormal_lab_results (−2.12)
Conjunctivitis	right_eye.pupil (+1.89), right_eye.other (+1.61), drug_orders (+1.59)	left_eye.vitreous (−1.10)

Table 15: Per-cohort schema-presence coefficients from the 39-field logistic regression, after the five structured diagnosis-presence flags are removed. The last three rows are the non-target control cohorts. A positive coefficient means that *populating* the field predicts the cohort; the strongest negative field is the one whose presence most argues against it, and for several cohorts it exceeds any positive coefficient in magnitude. No diagnostic field appears anywhere in the table: what separates the cohorts is which parts of the examination, order and per-eye documentation were filled in.

Removing the five diagnosis flags explicitly costs 0.020 macro-AUROC (0.8535 to 0.8337); restricting instead to the 37 fields that are non-constant on the training split leaves the result unchanged (0.8540). In the 39-field model the largest-magnitude coefficients are consistently examination-order, drug-order, laboratory-result, surgery-order and per-eye anatomical fields rather than anything diagnostic. Table 15 gives the three strongest positive presence fields per cohort together with the strongest negative one: presence of an ordered examination name is decisive for several cohorts, presence of surgery and laboratory orders separates the retinal-disease cohorts, and presence of per-eye free-text observation fields separates the non-target controls. The channel is

reading which parts of the clinical workflow were carried out and written down — the missingness shortcut of Section 5.4 viewed from the coefficient side.

F.5 Physician-decision sub-fields

The physician-decision channel bundles five order fields. Probing each in isolation, with the same frozen encoder, linear head and three seeds, shows the disease signal is not spread evenly across them (Table 14). Drug orders alone recover most of the combined channel’s floor, while laboratory and surgery orders are close to the range where seed variance dominates.

This complements the split reported in Section 5.2 between decisions (0.832) and raw lab-

oratory results (0.583): the recoverable signal sits in the therapeutic decision itself, and most sharply in what was prescribed.

modal fusion to avoid these shortcuts is the most direct empirical question this work opens.

F.6 Site and cohort cross-tabulation

To test whether a leave-one-site-out or hospital-balanced evaluation could be run on this corpus, we cross-tabulated all 5,904 visits by site and cohort. The corpus spans 120 distinct sites. Requiring at least 30 visits per cohort for a held-out site to give a stable eight-way macro-AUROC, no site qualifies: the maximum number of cohorts reaching that support at any single site is three of eight, and while one site does contain all eight cohorts, its smallest cohort holds three visits. The site distribution is strongly cohort-dependent by construction — several sites contribute only refractive-error controls, others only uveitis and retinal disease — because the corpus was assembled from disease-stratified exports. Site identity and cohort are therefore entangled at the level of the export, not merely correlated, and the automated feasibility gate in the released pipeline correctly skips the leave-one-site-out step rather than reporting an unstable number. The consequence for interpretation is stated in the Limitations: the channel floors we report are, if anything, an upper bound on how much site-independent clinical signal remains, and separating documentation style from clinical content requires replicating CRAFT on an independent corpus.

G Future Work

CRAFT functions in this paper as an *ex-post* audit instrument: it measures the disease-label signal each rendered channel carries, but does not act on those measurements during training. The channel decomposition itself naturally suggests an *ex-ante* use. Each channel — text-side narrative, schema, physician-decision, numeric measurements, and, when images are reintroduced, each imaging modality — can be treated as a unit of fusion in its own right rather than concatenated into an undifferentiated input. Under this view, a multimodal architecture is parameterized by a task-conditional, channel-level gate, and the channel-floor report becomes an *a priori* signal that can regularize that gate: channels whose stand-alone recoverability is disproportionately high are exactly the ones whose contribution one would want to constrain at training time. Whether channel-level priors derived from CRAFT can constrain multi-