Operationalizing Annotation Evidence into Guidelines for LLM Annotators

Anonymous ACL submission

Abstract

Annotation guidelines encode the operational decisions behind reproducible information extraction datasets, yet reusable guidelines are often missing even when schemas and expert annotations exist. We study this setting in LLM-based NER over clinical radiology reports. Annotation guidelines, which are developed at the early stage of annotation projects, expose useful policy, but whether that policy translates into correct annotation behavior depends on how operationally each annotator can use it: higher-capability LLM annotators repurpose mature human guidelines competitively, while lower-capability annotators benefit more from guidelines generated from source expert annotation evidence that make the intended annotation policy more explicit. Evidence-gated moderation on a small target-domain pilot further closes this capability-dependent alignment gap, but gains are non-monotonic: mid-sized rule compositions work best on the target corpus and also transfer to external radiology subsets. These results suggest that annotation guidelines can be generated, repurposed, measured, and selectively refined as evidence-driven operational specifications that make annotation policy usable by LLM annotators.

1 Introduction

Annotation guidelines specify the operational policy of information extraction datasets. They determine span boundaries, inclusion criteria, modifier handling, and label assignment, all of which directly affect annotation consistency and benchmark validity (Artstein and Poesio, 2008; Klie et al., 2024). The importance of these decisions becomes clear in NER tasks, where semantically plausible alternatives can still be counted as errors when they violate project-specific annotation policy. For example, the same phrase may be treated as an entity, a modifier, an exclusion, or part of a larger expression depending on the dataset's operational conventions.

LLMs are increasingly used as annotators, but their reliability depends on whether that policy is made operational. Prior work shows that LLMs can provide useful labels in low-resource settings (Wang et al., 2021; Ding et al., 2023; Gilardi et al., 2023), but IE tasks remain difficult when labels depend on dataset-specific annotation conventions rather than general domain meaning (Jimenez Gutierrez et al., 2022; Fonseca and Cohen, 2024). Guideline-driven approaches respond to this problem by treating annotation rules as computational objects that can be supplied to, learned by, or refined for models (Pang et al., 2023; Sainz et al., 2024; Qi et al., 2024). Recent work further suggests that existing human annotation guidelines can be revised or repurposed for LLM annotation, while effective guideline use may depend on the annotator model's ability to operationalize those guidelines (Bibal et al., 2025; Kim et al., 2026).

These findings leave three practical gaps. First, early-stage guideline writing is itself an annotation activity: human annotators often begin with public schema documentation, annotate pilot cases, and gradually materialize implicit conventions as written guidance. Although LLMs can be used to generate or edit guidelines, it remains unclear how closely they can approximate this early-stage guideline-writing process. Second, human-written guidelines are valuable because they preserve costly expert annotation decisions, but their direct transfer to LLM annotators may be model-dependent. We therefore ask whether the same guideline source fits different contemporary LLM annotators in the same way. Guidelines that are sufficiently operational for one LLM may still be underspecified, over-constraining, or poorly matched for another. Inducing new guidelines from annotation evidence can make policy more explicit, but such evidence may require additional expert labeling effort. Third, if neither public documentation nor human guidelines fully align an LLM annota-

tor, we need to know whether annotation evidence can make the intended policy more explicit, and whether remaining errors can be reduced through moderation.

We study these questions in LLM-based clinical NER using a design that separates three roles: a target corpus with newly created annotations and no reusable guideline, source expert annotations for evidence-derived guideline construction, and external radiology subsets for transfer evaluation. This reflects the practical regime we target: unlabeled reports may be plentiful, but target-domain expert annotations and reusable operational guidelines are scarce. This design lets us compare public documentation, staged human guidelines, early-stage LLM-written guidelines, and source-evidence-derived guidelines, then test whether pilot-moderated refinements generalize beyond the target corpus.

We ground the study in three empirical questions. **Q1**: Can LLMs approximate the early-stage human guideline-writing process when given similar public schema documentation and pilot-style feedback? **Q2**: When does human-guideline reuse transfer to LLM annotators, and when does annotation evidence yield a more operational guideline? **Q3**: Given an evidence-derived guideline, can multi-LLM moderation reduce residual alignment errors, and do accepted local rules compose monotonically as more refinements are appended?

Our contributions are threefold. **(1) Empirical comparison**: We compare six guidance conditions—from schema-only prompting to staged human and evidence-derived guidelines—across multiple LLM annotators, showing that early-stage guidance sources do not have a single model-independent ordering. **(2) Evidence-derived operationalization**: We compare reusable human guidelines with guidelines induced from source expert annotations, clarifying when costly annotation evidence adds value for LLM annotators. **(3) Evidence-gated moderation**: We design and evaluate a multi-annotator-style moderation procedure that converts pilot discrepancies into gated local rules and tests how accepted rules compose on held-out and external corpora.

2 Related Work

Annotation guidelines in biomedical and clinical IE. Biomedical IE resources such as NCBI Disease, BC5CDR, CRAFT, and BioRED illustrate that annotation quality depends on more than label inventories. It depends on operational conventions for span scope, modifier treatment, and difficult edge cases (Doğan et al., 2014; Li et al., 2016; Bada et al., 2010; Luo et al., 2022). Prior work on agreement and annotation quality management likewise treats guidelines as infrastructure for consistent, reusable labeling (Artstein and Poesio, 2008; Klie et al., 2024). This perspective is especially relevant for exact-match NER evaluation, where different yet semantically reasonable span choices count as errors. Clinical radiology annotation exhibits the same issue: RadGraph supplies expert annotations over MIMIC-CXR and CheXpert reports, while IU X-ray is a separate public chest X-ray corpus without pre-existing RadGraph-style annotations or reusable annotation guidelines (Jain et al., 2021; Demner-Fushman et al., 2016).

LLMs as annotators. LLMs have shown promise as scalable annotators across classification and structured labeling tasks (Wang et al., 2021; Ding et al., 2023; Gilardi et al., 2023; He et al., 2024). Yet IE exposes a persistent weakness: models can understand the text while still violating task-specific boundary and label conventions (Jimenez Gutierrez et al., 2022; Fonseca and Cohen, 2024). This distinction matters because expert annotations often encode historical project decisions rather than universally intuitive label assignments. For LLM-based annotation, the challenge is therefore not only whether a model can identify clinically salient content, but whether it can reproduce the dataset's operational policy consistently. Guideline following may itself vary across annotator models.

Guideline-driven and guideline-refinement methods. Recent IE work makes annotation guidance explicit. Pang et al. (2023) propose a guideline learning framework that synthesizes per-error rules from labeled training instances, building a growing rule bank rather than a structured operational document. GoLLIE trains models to follow detailed task guidelines in zero-shot extraction (Sainz et al., 2024). Related alignment work adapts LLMs to structured IE outputs (Qi et al., 2024). Other studies use LLMs to revise or repurpose existing annotation manuals (Bibal et al., 2025). Our work connects these threads by comparing public schema documentation, LLM-written initial guidelines, staged human-written guidelines, and evidence-derived guidelines, then refining residual

errors through pilot-based moderation.

3 Methodology

3.1 Problem Formulation

Figure 1 provides an overview. Let $\mathcal{S}$ denote the entity schema: labels and short descriptions (Appendix Table 7). A guideline $\mathcal{G}$ is any document that operationalizes $\mathcal{S}$ for annotation; the ideal operational guideline $\mathcal{G}^*$ is not available as a reusable public artifact. Given report text x, guideline $\mathcal{G}$, and LLM annotator θ, the annotator produces $\hat{y}_\theta(x; \mathcal{G})$. Gold labels are span-and-type tuples $y^*(x) = \{(s_i, e_i, t_i)\}_{i=1}^{n_x}$, where i indexes gold entities in x, s_i and e_i are offsets, and t_i is the entity type. We evaluate exact span-and-type micro-F1.

We distinguish three kinds of guidance. Public paper documentation extends $\mathcal{S}$ with schema-related notes and illustrative examples, but no source annotation evidence. Human V1 and V2 document decisions made during staged IU annotation. Evidence-derived generation uses a guideline generator q to produce $\mathcal{G}_{\text{init}} = q(\mathcal{S}, E)$—the initial evidence-derived guideline before pilot moderation—from compact evidence E distilled from source-corpus gold annotations D_{src}. D_{src} is used only to construct evidence and is distinct from target gold labels $y^*(x)$. The moderation experiment starts from $\mathcal{G}_{\text{init}}$ and asks whether residual errors can be reduced by pilot-based refinement.

3.2 Guideline Generation Resources and Format

We evaluate six main guideline/documentation conditions: schema only, paper documentation, LLM guideline V1, Human V1, Human V2, and $\mathcal{G}_{\text{init}}$. LLM guideline V1 denotes the early-stage guideline variants that each LLM annotator drafts from public documentation plus pilot-style feedback, without source expert annotations. Thus, each annotator model writes and uses its own LLM guideline variant. Human V1 and V2 come from staged IU annotation; and $\mathcal{G}_{\text{init}}$ uses compact MIMIC train/dev evidence, including mention strings, positive examples, boundary-sensitive cases, and cross-label contrasts. Generated guidelines follow a structured markdown format with task scope, global rules, entity-specific sections, boundary and exclusion rules, and examples. Appendix Table 8 reports $\mathcal{G}_{\text{schema}}$, a format-control diagnostic generated in the same structured format as $\mathcal{G}_{\text{init}}$ but without source annotation evidence.

3.3 Pilot Moderation and Patch Composition

Starting from $\mathcal{G}_{\text{init}}$, moderation tests whether pilot-derived local revisions improve annotation behavior and compose when appended to a generated guideline. We refine $\mathcal{G}_{\text{init}}$ on the 12-report pilot D_{pilot} through three stages: (i) discrepancy extraction and grouping, (ii) patch candidate generation, and (iii) evidence-gated acceptance.

We call this multi-annotator-style moderation because it mirrors human-style annotation moderation: multiple LLM annotators label the same pilot reports, recurring disagreements with gold annotations identify candidate policy gaps, and a fixed editor model proposes local revisions. The rationale is that errors recurring across annotators are more likely to reflect underspecified guideline policy than idiosyncratic model failures. A *patch* is a local guideline amendment targeting one recurring discrepancy pattern with a rule, examples, and an over-correction guard. We distinguish source evidence E for generating $\mathcal{G}_{\text{init}}$ from pilot acceptance evidence for moderation. Evidence gates accept a patch only when pilot performance supports the update without materially harming individual annotators or entity types. They separate improvement, regression, and validity concerns so that a patch fixes a targeted error without schema drift or over-correction (Appendix A.3).

Discrepancies (false negatives, false positives, boundary mismatches, and label mismatches) are grouped by error category and label pair, then clustered by the fixed LLM editor into recurring error groups. Groups supported by at least two documents or two models are eligible for patch generation and prioritized by supporting pilot cases, documents, and models. Accepted patches are appended as auditable refinement sections rather than used to regenerate the full guideline (examples in Appendix Tables 11–12). Let $P = \{p_1, \ldots, p_m\}$ be the accepted patch set and $r(k)$ the index of the k-th patch under ranking r. Starting from $\hat{\mathcal{G}}_0^{(r)} = \mathcal{G}_{\text{init}}$,

$$\hat{\mathcal{G}}_k^{(r)} = \hat{\mathcal{G}}_{k-1}^{(r)} \oplus p_{r(k)}.$$

We evaluate $k \in \{1, 3, 5, 10, 15, 20, 21\}$ under pilot-delta ranking (single-patch pilot F1 gain) and sequential ranking (moderation order).

This design separates source annotations, which provide broad evidence for constructing an initial operational guideline, from the target pilot, which

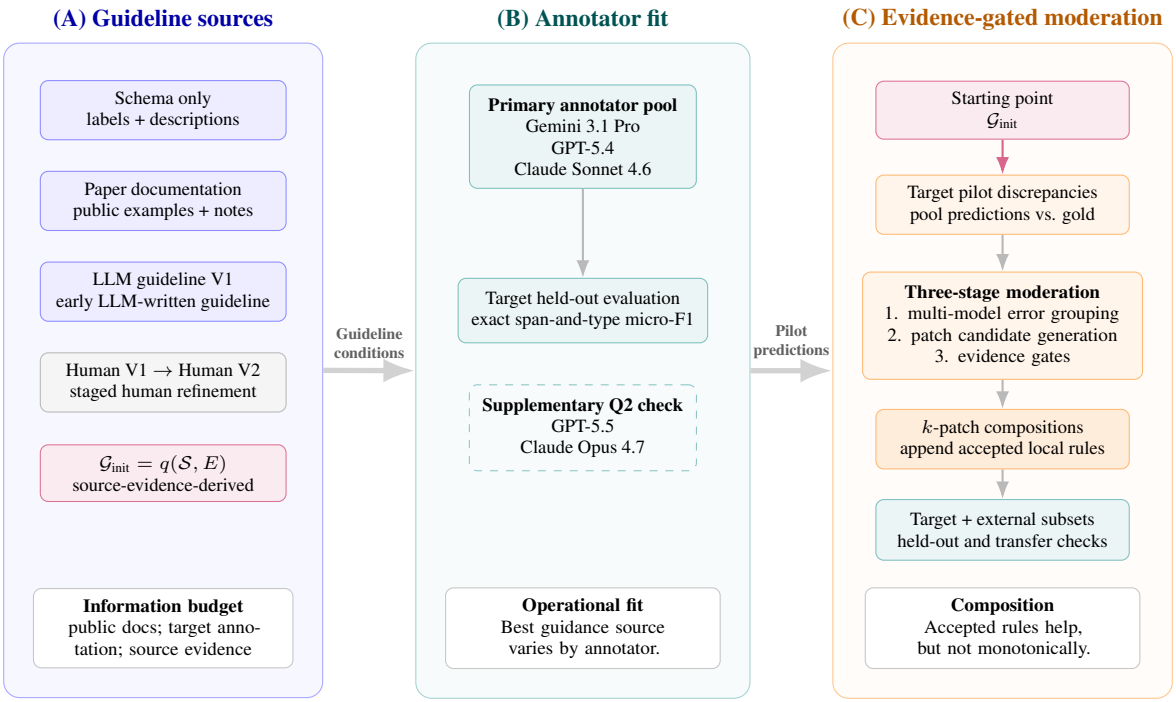

Figure 1: Overview of the study. (A) We compare six guideline sources with different information budgets. (B) We evaluate their fit to a mixed-capability LLM annotator pool, with supplementary high-capability annotators for the Q2 transfer check. (C) We refine $\mathcal{G}_{\text{init}}$ through evidence-gated moderation and test transfer of composed k-patch guidelines.

diagnoses corpus-specific omissions. Moderation therefore adapts guideline policy without using held-out target reports for optimization.

4 Experimental Setup

4.1 Datasets

All corpora share the same RadGraph annotation schema. MIMIC train/dev supplies source evidence for $\mathcal{G}_{\text{init}}$ construction; the IU pilot supplies moderation cases; IU held-out reports supply the primary test; and MIMIC/CheXpert test subsets supply external transfer checks (Table 1). Source annotations are not used as IU evaluation targets.

From the larger IU X-ray collection, the target-domain component uses two subsets annotated in the RadGraph style by our human expert annotator using the PubAnnotation framework (Kim et al., 2019): a 12-report pilot for moderation and a disjoint 38-report held-out test set.[1] The 12-report pilot corresponds to the first-round IU annotation batch available for moderation, rather than a hyperparameter tuned for this study. Human guidelines were produced at each annotation stage, V1 after the 12-report pilot and V2 after the full 50-report annotation. External transfer uses RadGraph MIMIC and CheXpert subsets, each annotated by two expert labelers; cross-corpus metrics average across labelers, with per-labeler significance tests in Appendix Table 15.

Dataset	Size	Role
RadGraph MIMIC train/dev	425 + 75 reports	Source evidence
IU X-ray pilot	12 reports	Moderation / refinement
IU X-ray held-out	38 reports	Primary test set
RadGraph MIMIC	50 reports × 2 labelers	External transfer
RadGraph CheXpert	50 reports × 2 labelers	External transfer

Table 1: Datasets used for guideline refinement and evaluation.

4.2 Models and Evidence

Gemini 3.1 Pro serves as the fixed guideline-writing and moderation editor: it writes generated guideline variants, clusters pilot discrepancies, and converts eligible groups into candidate rules, examples, exceptions, and over-correction guards. We use one editor model to keep generation and

[1] Our IU X-ray annotations, generated guidelines, accepted patch artifacts, and evaluation code will be released upon acceptance.

moderation comparable across conditions. Every generated guideline and accepted composition is evaluated with multiple LLM annotators, including non-Gemini models.

The primary annotator pool consists of three contemporary LLM annotators, Gemini 3.1 Pro, GPT-5.4, and Claude Sonnet 4.6, chosen to span different capability levels and model families. The same three models produce pilot predictions for discrepancy-based moderation and held-out predictions for evaluation, so guideline quality is judged by transfer across annotators rather than by the strongest model alone.

To test whether human-guideline transfer changes for high-capability annotators, we additionally evaluate selected conditions with GPT-5.5 and Claude Opus 4.7. These supplementary annotators do not participate in moderation and are used only for the Q2 diagnostic (Section 5.2), keeping Q3 scoped to the primary mixed-capability pool.

4.3 Conditions, Alignment, and Evaluation

The initial held-out comparison evaluates six conditions: schema only, paper documentation, LLM guideline V1, Human V1, Human V2, and $\mathcal{G}_{\text{init}}$.

All experiments use sentence-grounded outputs: reports are segmented, prompts request `(sentence_id, mention, type)`, and mentions are deterministically aligned back to offsets when uniquely matched within the indicated sentence, reducing offset hallucination (Xie et al., 2023; Semin et al., 2026) and repeated-string ambiguity. Significance uses paired document-level bootstrap resampling with 10,000 samples, one-sided $P(\Delta \leq 0)$.

5 Results

5.1 Early-stage Guideline Writing from Public Documentation (Q1)

Table 2 reports IU held-out performance for six baseline conditions, five of which use no source gold annotation evidence. Bare schema prompting is weak (mean F1 0.217), while paper documentation more than doubles mean F1 to 0.440, indicating that schema-related examples and explanatory documentation carry useful annotation policy. LLM guideline V1 is comparable on average (0.421) but does not consistently improve over paper documentation.

The ordering among paper documentation, LLM guideline V1, and Human V1 is not uniform across annotators: GPT-5.4 performs best with paper documentation, Claude Sonnet 4.6 with LLM guideline V1, and Gemini with Human V1. Precision/recall breakdowns (Appendix Table 10) show that paper documentation preserves substantially more recall for GPT-5.4 than the longer early-stage guidelines. Together, these results suggest that LLMs can extract useful conventions from public documentation and pilot-style feedback, but early guidance sources do not have a single model-independent ordering. Guideline usefulness depends not only on written policy, but on how a given LLM annotator operationalizes it.

Condition	Gemini	GPT-5.4	Claude	Mean
Schema only	0.342	0.161	0.147	0.217
Paper documentation	0.717	0.374	0.230	0.440
LLM guideline V1	0.689	0.224	0.351	0.421
Human guideline V1	0.784	0.278	0.326	0.463
Human guideline V2	0.828	0.455	0.491	0.591
$\mathcal{G}_{\text{init}}$	**0.834**	**0.744**	**0.677**	**0.752**

Table 2: IU X-ray held-out exact span-and-type micro-F1 across the primary LLM annotator pool. Mean averages Gemini, GPT-5.4, and Claude Sonnet 4.6.

5.2 Human Guideline Transfer Is Model-Dependent (Q2)

The human-written guideline conditions provide a staged diagnostic for documentation written during expert annotation. Human V1 was written after the 12-report pilot, while Human V2 was written after all 50 IU reports, including the held-out partition. V2 therefore has direct target-domain annotation experience, whereas $\mathcal{G}_{\text{init}}$ is induced from larger but external MIMIC evidence. V2 is best read as a human-moderated endpoint. Because both human guidelines were written for human annotators rather than as LLM prompts, we evaluate direct transfer to LLM annotators rather than intrinsic guideline quality.

The V1-to-V2 improvement (mean F1: 0.463 $\rightarrow$ 0.591) suggests that additional annotation experience makes human guidelines more useful to LLM annotators. Yet $\mathcal{G}_{\text{init}}$ is higher for all three primary annotators (mean 0.752), showing that source expert annotation evidence can make policy more operational when public or human documentation remains underspecified. The $\mathcal{G}_{\text{schema}}$ control, which uses the same structured format as $\mathcal{G}_{\text{init}}$ but no source evidence, is much lower (mean IU F1 0.292; Appendix Table 8), supporting the role of evidence content rather than format alone.

The transfer gap appears to follow a capability

Annotator	Human V2	$\mathcal{G}_{\text{init}}$	Δ
GPT-5.5	0.830	0.826	−0.004
Claude Opus 4.7	0.677	0.682	+0.005
Mean	0.753	0.754	+0.001

Table 3: Supplementary high-capability annotator check on IU held-out. GPT-5.5 and Claude Opus 4.7 are used only for the Q2 transfer diagnostic and do not participate in moderation.

k	Gemini	GPT-5.4	Claude	Mean
$\mathcal{G}_{\text{init}}$	0.834	0.744	0.677	0.752
1	0.840	0.788	0.766	0.798†
3	0.847	0.786	0.777	0.804†
5	0.851	0.803	0.792	0.815†
10	0.870	0.815	0.814	**0.833**†
15	0.860	0.814	0.816	0.830†
20	0.852	0.810	0.812	0.825†
21	0.855	0.812	0.809	0.825†

Table 4: Pilot-delta-ranked cumulative patch accumulation on IU held-out. The sequential-order sweep counterpart appears in Appendix Table 9. †: model-mean significantly improves over $\mathcal{G}_{\text{init}}$ ($p<0.05$, paired bootstrap, one-sided).

gradient. Human V2 and $\mathcal{G}_{\text{init}}$ are comparable for Gemini (0.828 vs. 0.834; $p = 0.31$), and the supplementary high-capability check shows the same pattern for GPT-5.5 and Claude Opus 4.7 (model mean 0.753 vs. 0.754; $p = 0.48$; Table 3). In contrast, $\mathcal{G}_{\text{init}}$ is much higher for GPT-5.4 and Claude Sonnet 4.6 ($\mathcal{G}_{\text{init}}$ − Human V2: +0.290 and +0.186; both $p<0.0001$), yielding a primary model-mean gap of +0.160. Thus, human-written guidelines remain valuable reusable artifacts, but direct repurposing is annotator-dependent. The comparison measures LLM transfer under our prompting format, not intrinsic guideline quality for human annotators.

These Q2 findings motivate moderation: although $\mathcal{G}_{\text{init}}$ makes many conventions explicit, residual errors remain across all three primary annotators. Evidence-gated moderation converts recurring discrepancies in the mixed-capability pilot pool into structured local rules, and the k-sweep tests whether these additional operational details continue to help when composed.

5.3 Moderation Improves Performance, but Gains Are Non-Monotonic (Q3)

Evidence-gated moderation of $\mathcal{G}_{\text{init}}$ yields 21 individually accepted patches (Appendix Table 16). Because these patches are discovered from recurrent discrepancies across the primary annotator pool, they test whether refinement closes residual alignment gaps rather than merely helping the strongest annotator. Since the same three models serve as pilot and held-out annotators, external transfer (Section 5.4) is the key generalization check.

To test whether operational detail composes monotonically, we evaluate two rankings over the same patch pool: *pilot-delta* ranking by single-patch pilot F1 gain and *sequential* moderation order. Table 4 reports pilot-delta accumulation; Appendix Table 9 reports the sequential counterpart. The key question is whether accepted patches continue to compose beneficially as more rules are added.

All pilot-delta compositions from k=1 through k=21 significantly improve over $\mathcal{G}_{\text{init}}$ ($p<0.05$), but gains are non-monotonic. Mean F1 rises from 0.798 at k=1 to 0.833 at k=10 (+0.081 over $\mathcal{G}_{\text{init}}$), then declines to 0.825 at k=20–21. The k=10 gains are largest for the lower-capability annotators (Claude: +0.137, GPT-5.4: +0.071) and smallest for Gemini (+0.036), mirroring the capability-dependent transfer gap in Q2.

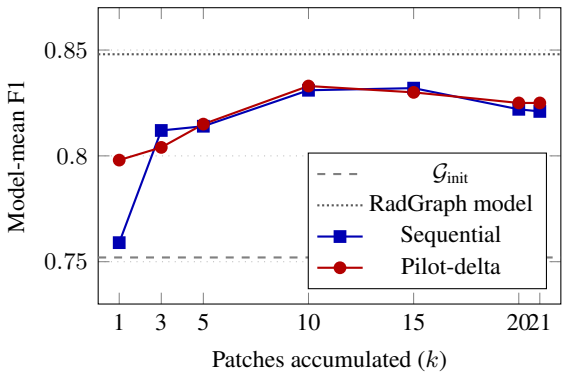

Figure 2: Model-mean F1 under both accumulation strategies over the same 21-patch pool, with $\mathcal{G}_{\text{init}}$ (0.752) as baseline. The dotted line marks a supervised RadGraph model trained on MIMIC as a calibration reference. Both curves share the same non-monotonic shape: rising to a peak at k=10–15, then declining modestly at k=20–21.

The sequential sweep shows the same qualitative pattern, peaking near k=15 (0.832) and declining at k=20–21 (Figure 2). Because both strategies draw from the same patch pool, this is not independent replication. Nevertheless, the shared shape suggests that the non-monotonicity is not an artifact of one ordering. For calibration, a supervised RadGraph model trained on MIMIC reaches F1 0.848 on IU after output remapping. The k=10 model mean reaches 0.833 using only a 12-report target pilot and no target-domain model training.

Data	Model	$\mathcal{G}_{\text{init}}$	Pilot-delta k=10
MIMIC	Claude	0.825	0.885
MIMIC	Gemini	0.893	0.911
MIMIC	GPT-5.4	0.868	0.900
CheXpert	Claude	0.737	0.827
CheXpert	Gemini	0.856	0.849
CheXpert	GPT-5.4	0.778	0.844

Table 5: Cross-corpus transfer of the pilot-delta k=10 IU-refined guideline, averaged across two labeler references per corpus. Per-labeler significance tests appear in Appendix Table 15.

Patch	Bef.	Aft.	Reduction	ΔF1
Imaging procedure targets	19	6	68.4%	+0.018
Anatomical modifier splitting	7	3	57.1%	+0.069
Within normal limits decomp.	25	0	100.0%	+0.039

Table 6: Pilot target error counts before and after applying each of the three patches in the sequential k=3 composition. Target error reduction is a diagnostic outcome, not an acceptance criterion. All 21 accepted patches are listed in Appendix Table 16.

5.4 Pilot Refinements Transfer to External Corpora

We evaluate the pilot-delta k=10 composition, the peak pilot-delta composition in the IU analysis, on RadGraph MIMIC and CheXpert. Table 5 reports labeler-averaged F1. The guideline improves model-mean performance on both external corpora, mainly through GPT-5.4 and Claude. Gemini improves on MIMIC and is slightly lower than $\mathcal{G}_{\text{init}}$ on CheXpert. Because patches are discovered on IU rather than tuned on external test subsets, these gains suggest that moderation captures conventions that transfer beyond isolated pilot reports. MIMIC gains may partly reflect the MIMIC evidence used in $\mathcal{G}_{\text{init}}$. CheXpert provides the cleaner transfer test.

The transfer pattern mirrors IU while showing model-level heterogeneity; we therefore treat k=10 as a strong evidence-based composition, not a universally optimal patch count.

5.5 Why Accepted Patches Do Not Compose Monotonically

The k-sweep poses a puzzle: each of the 21 accepted patches was individually validated by the evidence gates, yet applying all 21 is not best under either accumulation strategy. Two observations sharpen the question. First, patch rank alone does not determine composition quality: at k=3, the pilot-delta and sequential compositions share only one patch, and the higher-ranked pilot-delta set does not outperform the sequential set on held-out IU (0.804 vs. 0.812). Second, the patches that compose best are qualitatively distinct from those that add less or harm performance; examining the early patches concretely motivates the mechanisms proposed below.

Table 6 shows target error counts for the three patches in the sequential k=3 composition, which illustrate what moderation produces at its most effective. *Imaging procedure targets* corrects cases where models miss broad anatomical regions (e.g., *chest*, *abdomen*) inside imaging recommendations. *Anatomical modifier splitting* addresses the span-granularity disagreements documented in the original RadGraph paper: models predict compound spans (e.g., *right middle lobe*) while gold annotates each constituent word separately. *Within-normal-limits decomposition* resolves a fixed-expression boundary in which gold applies the single-word tagging convention to all three constituents of "within normal limits." A paired accepted/rejected example appears in Appendix Table 13.

These patches address broad, corpus-wide conventions that affect many held-out reports. Two candidate mechanisms may explain why composing the full pool does not maintain these gains. The first is a qualitative shift in patch scope. The top-ranked patches in pilot-delta ordering address broad, corpus-wide conventions—span granularity, fixed-expression decompositions, spatial modifiers—that affect many held-out reports. The patches ranked 11–21 are narrower: their individual pilot gains are all below $+0.030$, they target finer label-boundary decisions, and six of the eleven address modifier categories that cumulatively expand OBS-DP coverage beyond what any single regression gate would reject. Each individual expansion passes the entity-type gate; their aggregate effect may erode precision in ways that patch-by-patch evaluation cannot anticipate.

The second mechanism is direct rule conflict. Two accepted patches address the same surface form—de-identified tokens (XXXX, used when patient-identifiable text is masked)—but assign different labels: one rule instructs annotators to tag XXXX as ANAT when it replaces an anatomical location; a second instructs them to tag XXXX as OBS when it replaces a clinical observation. Each patch individually reduces its target pilot errors and passes all regression gates. When both are

present, however, the annotator receives contradictory guidance for the same token form. In the pilot-delta ordering, the ANAT patch enters at rank 7—before the k=10 peak—while the contradictory OBS patch enters at rank 14; both rules are active in all compositions from k=15 onward. In the sequential ordering, the OBS patch enters at position 8 and the ANAT patch at position 18, so both are present from k=18 onward, consistent with the sequential decline emerging at k=20–21. The per-model pattern is also informative: Gemini, the highest-performing annotator, shows the largest absolute drop from k=10 to k=21 (-0.015), while lower-capability annotators remain closer to their k=10 performance (-0.003 and -0.005). A model already near ceiling in those annotation areas may be most sensitive to conflicting guidance, whereas lower-capability annotators have more residual error and thus less to lose from rule conflicts they may partially ignore. Together, these patterns suggest a general failure mode: independence-based acceptance gates cannot fully anticipate how rules interact under joint application, and the composition ceiling may be partly determined by conflicts that only become visible when patches are evaluated as a composed whole.

6 Discussion

The central result is that annotation guidelines function as operational policy, not merely as containers for more instructions. The Q1 results show that adding policy does not create a model-independent ordering. Compact public documentation, early LLM-written guidelines, and Human V1 each fit different annotators best. The implication is that guideline design should be evaluated as annotator-specific operational fit, not only as document completeness.

The human-guideline comparison refines, rather than rejects, guideline reuse. Human guidelines preserve costly expert annotation decisions, and high-capability LLM annotators can use Human V2 competitively. At the same time, lower-capability annotators fall substantially short under direct transfer and benefit more from evidence-derived operationalization. Thus, the practical question is not whether human guidelines are valuable, but when direct reuse is sufficient and when additional annotation evidence is needed to make the policy operational for a given LLM annotator.

Motivated by this unevenness, the main moderation experiment treats residual errors under $\mathcal{G}_{\text{init}}$ as a multi-annotator alignment problem: recurring errors across LLM annotators reveal where the evidence-derived guideline is not yet operational enough for all annotators. Accepted patches clarify underspecified decisions and selected compositions generalize beyond the pilot. Rejected candidates show the complementary risk: a locally correct rule can reduce global F1 when it over-specifies decisions that LLMs apply beyond the intended scope, triggering entity-type or per-model regressions even when its target error group is eliminated (Appendix Table 13).

Across the three questions, a common mechanism emerges: alignment failures arise when a guideline's design assumptions do not match the annotator using it. Compact documentation may preserve recall while leaving policy implicit; human and LLM-written guidelines add policy but also introduce structure or emphasis that different LLMs interpret differently. Evidence-derived guidelines make many conventions explicit but still leave residual ambiguities. Patch composition adds another layer: individual rule validity does not guarantee joint usability. The qualitative scope shift at higher k and the de-identified-token conflict suggest that composing refinements into a non-conflicting guideline is itself an empirical problem.

7 Conclusion

We study LLM annotation when an operational guideline is missing but public schema documentation, source expert annotations, and a small target-domain pilot are available. In RadGraph-style radiology NER, early LLM-written and human-written guidelines expose useful policy, but their value depends on how each LLM annotator operationalizes the instructions. Human-written guidelines remain valuable reusable artifacts, yet direct transfer is model-dependent. Evidence-derived guidelines and evidence-gated moderation are helpful for annotators that need more explicit operationalization. Moderation improves held-out performance and transfers to MIMIC and CheXpert, but gains are non-monotonic. Overall, annotation guidelines should be treated as measurable operational specifications whose local revisions must be evaluated in composition within LLM annotation pipelines.

Limitations

Our primary target annotation set is small: 50 IU X-ray reports, split into a 12-report pilot and a 38-report held-out set. This scale matches the setting we study—guideline operationalization when only a small target-domain pilot is available—but it limits statistical power and does not establish a corpus-level benchmark for IU X-ray. Paired bootstrap tests measure stability under document resampling, not general validity under different pilot draws, larger target annotations, or future model releases. External MIMIC and CheXpert transfer checks partially mitigate this concern, but do not remove the need for larger target-domain evaluation.

The study is also limited to English chest X-ray NER under the RadGraph schema. Transfer to relation extraction, other clinical domains, other languages, or schemas with different ambiguity structures remains open. Our results should therefore be read as evidence about guideline operationalization and pilot-based refinement in this controlled radiology IE setting, not as a general claim about all biomedical annotation tasks.

The moderation design may reflect the annotator and editor pool. The same three primary LLM annotators provide pilot predictions and held-out evaluations, so accepted patches may partially track this pool's error tendencies. Gemini 3.1 Pro also serves as the fixed guideline-writing/moderation editor and one evaluated annotator, creating a possible self-alignment bias. Improvements for GPT-5.4 and Claude Sonnet 4.6, and the smaller Gemini gain at $k{=}10$, suggest that results are not only Gemini self-fit. Nevertheless, future work should separate editor and annotator roles and evaluate additional annotator pools.

Several design choices remain heuristic. Discrepancy prioritization, stopping decisions, and evidence-gate thresholds are operational screening criteria rather than optimized hyperparameters; different thresholds could change which patches are accepted and how many should be composed. Human V2 and $\mathcal{G}_{\text{init}}$ also differ in both information source and document format, and we do not moderate Human V2 or measure human annotator performance under Human V1/V2. Matched-format controls, human-guideline moderation, and gate-threshold sensitivity analyses are important future work.

Finally, some interpretations should be read as diagnostic rather than causal. The supplementary GPT-5.5 and Claude Opus 4.7 checks support the capability-gradient interpretation but do not define a general model-capability scale. Our explanations for non-monotonic patch composition are also candidate mechanisms; attention dilution from append-style composition or pilot-specific over-correction may contribute as well. The strongest evidence remains the primary held-out improvements and their transfer to external corpora.

Ethics Statement

This work studies annotation-policy alignment for radiology reports using de-identified research datasets and does not make clinical predictions for patient care. RadGraph is a restricted-access PhysioNet resource used under the PhysioNet Credentialed Health Data License 1.5.0 and Data Use Agreement 1.5.0. All required credentialing, training, and data-use requirements were fulfilled. IU X-ray was used as a publicly available dataset under its release terms. We did not collect new patient data, redistribute restricted report text, or attempt to re-identify individuals. We rely on the source datasets' approved release procedures for access and consent conditions. We report only aggregate results and generic qualitative examples.

The IU X-ray annotations and human-written guidelines were produced by a biomedical annotation expert who was compensated at a standard hourly rate and later contributed to manuscript preparation. The annotator was not exposed to identifiable patient data and did not participate in LLM evaluation or result analysis. No crowdworkers or external unpaid annotators were recruited. A remaining risk is methodological: generated guidelines or LLM annotations may be over-trusted outside carefully evaluated settings. We therefore evaluate against expert annotations, distinguish primary from exploratory analyses, and avoid presenting generated guidelines as substitutes for domain-expert review in clinical deployment.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Michael Bada, Lawrence E. Hunter, Miriam Eckert, and Martha Palmer. 2010. An overview of the CRAFT concept annotation guidelines. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages

207–211, Uppsala, Sweden. Association for Computational Linguistics.

Adrien Bibal, Nathaniel Gerlek, Goran Muric, Elizabeth Boschee, Steven C. Fincke, Mike Ross, and Steven N. Minton. 2025. Automating annotation guideline improvements using LLMs: A case study. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 129–144, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Lion Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Marcio Fonseca and Shay Cohen. 2024. Can large language models follow concept annotation guidelines? a case study on scientific and financial domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8027–8042, Bangkok, Thailand. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *PhysioNet*. Version 1.0.0.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jin-Dong Kim, Yue Wang, Toyofumi Fujiwara, Shujiro Okuda, Tiffany J Callahan, and K Bretonnel Cohen. 2019. Open agile text mining for bioinformatics: the pubannotation ecosystem. *Bioinformatics*, 35(21):4372–4380.

Kon Woo Kim, Jin-Dong Kim, and Akiko Aizawa. 2026. Refining and reusing annotation guidelines for llm annotation. *Preprint*, arXiv:2605.20809.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: A resource for chemical disease relation extraction. *Database*, 2016:baw068.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. BioRED: A rich biomedical relation extraction dataset. In *Briefings in Bioinformatics*, volume 23, page bbac282.

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15372–15389, Singapore. Association for Computational Linguistics.

Yunjia Qi, Hao Peng, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2024. ADELIE: Aligning large language models on information extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Miami, Florida, USA. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*.

Danil Semin, Ondřej Dušek, and Zdeněk Kasner. 2026. Strategies for span labeling with large language models. *Preprint*, arXiv:2601.16946.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Transfer and Statistical Diagnostics

Bootstrap tests. We use paired document-level bootstrap because each comparison evaluates two conditions on the same reports. Resampling documents preserves report-level dependence among entity decisions, while pairing controls for document difficulty. The one-sided p value estimates how often the resampled F1 difference is non-positive; it therefore measures the stability of the observed paired improvement to document sampling variation, not general validity across unrelated tasks, pilot draws, or model releases.

A.2 Schema and Format-Control Diagnostics

Label	Short description used in schema-only prompts
ANAT	Anatomical entities, including body regions and anatomical structures.
OBS-DP	Definitely present observations.
OBS-DA	Definitely absent observations.
OBS-U	Uncertain observations.

Table 7: Compact excerpt of the entity schema $\mathcal{S}$ used in schema-only prompting. Full prompt resources will be released with the code and artifacts.

Annotator	$\mathcal{G}_{\text{schema}}$	$\mathcal{G}_{\text{init}}$
Gemini 3.1 Pro	0.410	0.834
GPT-5.4	0.245	0.744
Claude Sonnet 4.6	0.222	0.677
Mean	0.292	0.752

Table 8: IU held-out F1 for the schema-only generated guideline control. $\mathcal{G}_{\text{schema}}$ uses the same structured guideline format as $\mathcal{G}_{\text{init}}$ but omits source annotation evidence, supporting the interpretation that $\mathcal{G}_{\text{init}}$ gains come from evidence content rather than format alone.

A.3 Patch Composition and Evidence Gates

Evidence gate specifications. Patch acceptance applies a primary criterion, performance-based regression gates, and structured validity checks.

Primary criterion. A candidate patch is accepted if at least one of the following holds:

1. pilot model-mean F1 improves by $\geq +0.01$; or
2. the targeted error group is reduced by $\geq 20\%$ while model-mean F1 does not degrade by more than 0.005.

Performance-based regression gates. Even when the primary criterion is satisfied, a patch is rejected if any of the following held on the pilot set:

1. *Per-model regression*: any single model's F1 drops by more than 0.02.
2. *Per-label regression*: any entity-type F1 drops by more than 0.05.
3. *FP expansion*: false positives increase by more than 10% without a commensurate reduction in false negatives.

Structured validity checks. We also reject patches that are not scoped to one localized phenomenon, lack an explicit exception or overcorrection guard, or introduce labels outside the canonical schema.

Each gate addresses a qualitatively distinct failure mode: the per-model gate prevents mean-beneficial patches that substantially hurt one model; the per-label gate prevents entity-type regressions masked by aggregate metrics; the FP expansion gate prevents precision erosion without recall gain; and the structured validity checks prevent scope creep and schema drift.

The specific threshold values reflect a conservative design philosophy in which all gates are biased toward rejection. They are operational filters rather than optimized hyperparameters, and gate thresholds were fixed before held-out evaluation of accepted candidate guidelines.

k	Gemini	GPT-5.4	Claude	Mean
$\mathcal{G}_{\text{init}}$	0.834	0.744	0.677	0.752
1	0.828	0.745	0.704	0.759
3	0.847	0.798	0.792	$0.812^{\dagger}$
5	0.849	0.800	0.795	$0.814^{\dagger}$
10	0.862	0.820	0.811	$0.831^{\dagger *}$
15	0.859	0.821	0.815	$\mathbf{0.832}^{\dagger *}$
20	0.853	0.808	0.806	$0.822^{\dagger *}$
21	0.853	0.804	0.807	$0.821^{\dagger}$

Table 9: Sequential-order cumulative patch accumulation on IU held-out (38 reports); patch order from Table 16. †: model-mean significantly improves over $\mathcal{G}_{\text{init}}$ ($p<0.05$, paired bootstrap, one-sided). *: also significantly improves over $k=3$ ($p<0.05$).

Condition	Annotator	Precision	Recall	F1
Schema only	Gemini	0.470	0.269	0.342
	GPT-5.4	0.270	0.115	0.161
	Claude Sonnet 4.6	0.253	0.104	0.147
	Mean	0.331	0.162	0.217
Paper documentation	Gemini	0.757	0.681	0.717
	GPT-5.4	0.498	0.299	0.374
	Claude Sonnet 4.6	0.346	0.172	0.230
	Mean	0.534	0.384	0.440
LLM guideline V1	Gemini	0.738	0.646	0.689
	GPT-5.4	0.334	0.168	0.224
	Claude Sonnet 4.6	0.466	0.281	0.351
	Mean	0.513	0.365	0.421
Human guideline V1	Gemini	0.827	0.746	0.784
	GPT-5.4	0.452	0.201	0.278
	Claude Sonnet 4.6	0.462	0.252	0.326
	Mean	0.580	0.400	0.463
Human guideline V2	Gemini	0.867	0.793	0.828
	GPT-5.4	0.643	0.352	0.455
	Claude Sonnet 4.6	0.639	0.399	0.491
	Mean	0.716	0.515	0.591
$\mathcal{G}_{\text{init}}$	Gemini	0.876	0.796	0.834
	GPT-5.4	0.833	0.673	0.744
	Claude Sonnet 4.6	0.796	0.589	0.677
	Mean	0.835	0.686	0.752

Table 10: Precision, recall, and F1 for the IU held-out baseline families in Table 2. The breakdown shows that guideline detail and F1 do not vary monotonically: compact paper documentation preserves more recall than early-stage guideline variants for GPT-5.4, whereas Gemini benefits most from Human V1 among the early-stage conditions.

Moderation stage	Representative content for `within_normal_limits_components_013`
Error grouping	The discrepancy group collects cases where models omit *within* and *limits* in the idiom "within normal limits," while the gold convention annotates each constituent word separately as OBS-DP.
Pattern explanation	The model interprets these tokens as syntactic scaffolding rather than as guideline-relevant pieces of a fixed expression, producing recurrent false negatives.
Principle generation	Define a narrow fixed-expression rule that preserves the dataset's single-word tagging convention but adds explicit exceptions for non-idiomatic uses.
Patch realization	Insert a short OBS-DP positive rule, one annotated example sentence, and an over-correction warning into the generated guideline document.

Table 11: A compact end-to-end moderation trace from grouped discrepancy to guideline patch.

Abridged excerpt from $\mathcal{G}_{\text{init}}$

Global rule: Independent status modifiers. Descriptors of state, severity, change, or normality (e.g., *normal*, *clear*, *unchanged*) are annotated as independent observation spans.

OBS-DP positive rule. Extract words denoting a state of health or lack of disease (e.g., *clear*, *normal*, *unremarkable*).

Appended target-domain pilot refinement

Evidence group: `within_normal_limits_components_013`
Rule. For the fixed phrase "within normal limits," extract each component word as an independent OBS-DP span: `[within]`, `[normal]`, and `[limits]`.
Exception. Do not generalize this rule to spatial or boundary uses of *within* or *limits*, such as "within the right lung" or "upper limits."
Example. "The cardiac contour is within normal limits." → `[cardiac]` ANAT, `[contour]` ANAT, `[within]` OBS-DP, `[normal]` OBS-DP, `[limits]` OBS-DP.
Over-correction guard. Avoid predicting OBS-DP for *within* or *limits* outside this fixed idiom.

Table 12: Append-only guideline refinement excerpt. $\mathcal{G}_{\text{init}}$ gives a general normality rule; evidence-gated moderation appends a narrow rule with an exception, example, and over-correction guard.

Candidate	Target phenomenon	Observed local effect	Decision logic
Accepted: `within_normal_limits_components_013`	Recurring omission of *within* and *limits* in a fixed normality phrase	Target errors reduced from 25 to 0	Accepted because the patch removes the recurring pilot failure pattern while preserving overall stability.
Rejected: `findings_in_communication_context_014`	Model omission of clinical observations in administrative communication sentences (e.g., "was notified of [opacity]")	Target errors reduced from 11 to 0	Rejected because a label-type regression (max entity-type F1 drop -0.076) exceeds the -0.05 threshold, despite a positive pilot delta (+0.020) and complete target error elimination.

Table 13: Accepted versus rejected moderation examples. The rejected case shows why target-group correction alone is insufficient for guideline admission.

Dataset	Model	Prompt	Schema	$\mathcal{G}_{\text{schema}}$	$\mathcal{G}_{\text{init}}$	Pilot-delta k=10
MIMIC	Claude	0.183	0.247	0.340	0.825	0.885
MIMIC	Gemini	0.527	0.453	0.564	0.893	0.911
MIMIC	GPT-5.4	0.211	0.273	0.356	0.868	0.900
CheXpert	Claude	0.153	0.169	0.260	0.737	0.827
CheXpert	Gemini	0.426	0.372	0.482	0.856	0.849
CheXpert	GPT-5.4	0.158	0.165	0.289	0.778	0.844

Table 14: Cross-corpus prompt-family comparison, averaged across two labeler references per corpus. $\mathcal{G}_{\text{schema}}$ uses the same structured guideline format as $\mathcal{G}_{\text{init}}$ but omits source annotation evidence, isolating the contribution of evidence beyond format. The final column is the pilot-delta k=10 guideline used for transfer evaluation in the main text.

Comparison	Δ F1	95% CI	p
IU: Schema only $\rightarrow \mathcal{G}_{\text{init}}$ (model mean)	+0.535	[+0.502, +0.568]	< 0.0001
IU: Human V1 $\rightarrow \mathcal{G}_{\text{init}}$ (model mean)	+0.289	[+0.263, +0.314]	< 0.0001
IU: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (model mean)	+0.160	[+0.131, +0.190]	< 0.0001
IU: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (Gemini only)	+0.006	[-0.014, +0.026]	0.31
IU: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (Claude only)	+0.186	[+0.131, +0.244]	< 0.0001
IU: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (GPT-5.4 only)	+0.290	[+0.241, +0.337]	< 0.0001
IU supplementary: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (GPT-5.5 only)	-0.004	[-0.025, +0.020]	0.63
IU supplementary: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (Opus 4.7 only)	+0.005	[-0.023, +0.034]	0.37
IU supplementary: Human V2 $\rightarrow \mathcal{G}_{\text{init}}$ (GPT-5.5/Opus mean)	+0.001	[-0.020, +0.022]	0.48
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=1	+0.046	[+0.033, +0.060]	< 0.0001
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=3	+0.052	[+0.039, +0.066]	< 0.0001
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=5	+0.063	[+0.050, +0.078]	< 0.0001
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=10	+0.081	[+0.065, +0.098]	< 0.0001
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=15	+0.078	[+0.062, +0.095]	< 0.0001
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=20	+0.073	[+0.056, +0.091]	< 0.0001
IU: $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=21	+0.074	[+0.058, +0.091]	< 0.0001
MIMIC (L1): $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=10 (model mean)	+0.036	[+0.023, +0.052]	< 0.0001
MIMIC (L2): $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=10 (model mean)	+0.036	[+0.022, +0.052]	< 0.0001
CheXpert (L1): $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=10 (model mean)	+0.057	[+0.039, +0.076]	< 0.0001
CheXpert (L2): $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=10 (model mean)	+0.043	[+0.025, +0.061]	< 0.0001

Table 15: Selected paired-bootstrap tests (10,000 samples, one-sided) for the main IU baseline, human-guideline, supplementary high-capability annotator, composition, and transfer claims. "Pilot-delta k" rows refer to cumulative accumulation ordered by single-patch pilot F1 gain (Table 4). Cross-corpus rows report per-labeler model-mean tests for the primary $\mathcal{G}_{\text{init}} \rightarrow$ pilot-delta k=10 comparison. No multiple-comparison correction is applied; exploratory rows should be interpreted descriptively.

Seq.	Δ rk.	Patch theme	Pilot ΔF1
1	19	Anatomical targets of imaging procedures	+0.018
2	1	Lumping vs. splitting of anatomical modifiers	+0.069
3	4	Components of "within normal limits"	+0.039
4	3	Missed laterality and spatial modifiers	+0.039
5	15	Core anatomical structures labeled OBS-DP in gold	+0.025
6	17	Organ-based adjectival modifiers	+0.020
7	8	Spatial relation participles (*projecting, overlying*)	+0.034
8	14	Masked/de-identified tokens (XXXX) as OBS	+0.025
9	6	Adverbial, degree, and normalcy modifiers	+0.038
10	10	Observation noun phrase splitting	+0.032
11	18	Definitive finding with uncertain etiology	+0.019
12	2	Spatial subdivision nouns as ANAT	+0.042
13	13	Descriptive adjectival modifiers	+0.028
14	5	Locational modifiers labeled OBS-DP in gold	+0.038
15	21	Measurement values (*6 mm, 1.4 cm*)	+0.011
16	16	Temporal and trend modifiers	+0.023
17	20	Distribution modifier *scattered*	+0.016
18	7	De-identified tokens (XXXX) labeled as ANAT	+0.037
19	9	Abstract or non-clinical terms	+0.033
20	11	Abstract radiological features (OBS-DP vs. ANAT)	+0.028
21	12	Trailing punctuation in gold span	+0.028

Table 16: Accepted patch pool with both sequential moderation order and pilot-delta rank. Pilot ΔF1 is the independent single-patch mean F1 gain on the 12-report pilot; pilot-delta accumulation orders patches by this value. The horizontal rule separates the compact sequential k=3 case-study subset from the remaining accepted patches.