INTERIORALLOC: WORST-CASE INTERIOR ALLOCATION AND STATEMENT-LEVEL CREDIT FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation is trained and judged against one similarity summed over a whole document, while a clinician acts on the individual statements that document commits to. That summation cannot place a worst-case deficit inside an evidence source, credit a statement by its type, or keep a corrected report within what the generator can deliver, and a weight on the source leaves the within-source deficit untouched. We propose InteriorAlloc, built on one insight: robustness, credit and repair are three readings of one allocation, so a single update carries all three. The allocation is solved exactly inside each evidence source by convex duality, with the dual derivative at the returned root at most 2.03×10^{-13}; each typed statement is credited by exact leave-one-out under that allocation; and the action set is closed under repair, so every corrected report stays deliverable. Every reading is taken on four generated populations that carry the names of public corpora and are drawn from one operating point fixed by published figures, three at a fixed shift, none a measurement on the corpora themselves. Against three objective forms drawn from four published systems, InteriorAlloc reads best in 12 of 18 main-table columns, CheXbert F1 on every carrier among them, 0.449 against 0.4173 on MIMIC-CXR, and the stand-alone-credit or R2Gen form leads the other six. No repair inside the closure introduces a new error, while the three rewrites of a delivered report read new-error rates of 0.0867 to 0.1515, and on every carrier the closure delivers fewer pooled false positives than the same policy with the closure removed. Against a uniform within-source arm, 0.9372 of the true-positive mass on which the two arms' reports differ cancels in pooled counts, so the reallocation InteriorAlloc performs is visible only at the statement level a summed score cannot read.

1 INTRODUCTION

Chest radiograph report generation is trained (Miura et al., 2021) and judged (Banerjee & Lavie, 2005) by one number: the similarity between a generated report and a reference, summed over the whole document (Zhao et al., 2026b; Jing et al., 2026; Zhao et al., 2026a). This recipe is the current standard (Yu et al., 2023), and every baseline in this paper's tables follows it in one form or another. The summation returns a document-level scalar (Smit et al., 2020; Jain et al., 2021): it is the same number whatever the report commits to statement by statement, whatever evidence entry each statement stands on, and whether a corrected report is one the deployed generator could have emitted. A statement that records what was seen and a statement that asks for an action on the same entry are different clinical acts, and the summation assigns them the same credit. Those three are what decide whether a report is clinically right.

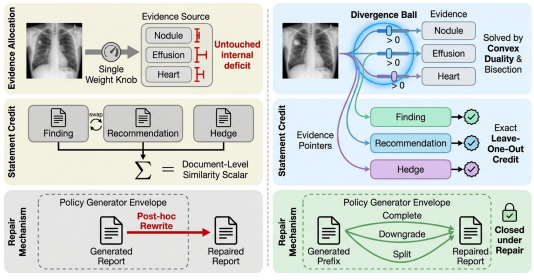


Figure 1: **Left:** one weight per source, one score per document and a rewrite after delivery. **Right:** InteriorAlloc allocates inside each source by convex duality, credits typed statements by leave-one-out and repairs within its own action set.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent report generators add evidential structure to what the model says: hierarchical evidential reasoning over candidate findings (Zhao et al., 2026a), differential visual semantic prompting (Tian et al., 2026), and radiologist-style chain-of-thought trained with reinforcement learning (Jing et al., 2026). These sharpen both the content of the report and its traceability back to the image. A second family asks whether one delivered clinical statement is supported, through medical hallucination benchmarks (Pal et al., 2023), fact-controlled diagnosis of hallucination in medical summarisation (BN et al., 2025), and counter-evidence verification (Zhou et al., 2026a). These establish that a single statement can be checked one at a time. A third family repairs a report after it is delivered, through rollout-augmented self-correction (Ding et al., 2026), applying a correction once a candidate already exists. Together, the three families have made the content, the checking and the correction of a radiology report progressively more explicit.

The three properties the summation cannot express are precisely the three these families work around. Robustness enters as a weight on an evidence source, which is one number per source: it moves the deficit lying between sources and leaves the part lying inside them where it was. Credit enters as a report-level score, which attaches to the document and moves for reasons unrelated to any statement's clinical content. Repair enters after delivery, which places the corrected form outside the class the policy can emit. The allocation over labels, the credit over documents and the repair outside the action set are thus one shortcoming rather than three: each is the same document-level summation surfacing at a different point in the loop.

To address these three, we propose InteriorAlloc, a training loop whose allocation is solved inside each evidence source, read on typed statements, and closed under repair. Its core insight inverts that reading: robustness, credit and repair are three readings of one allocation, so the solve that decides how mass is spread inside an evidence source also decides what a statement is worth and which corrected reports remain deliverable. One update carries all three: a divergence-ball allocation over the entries of each source solved by convex duality and bisection, an exact leave-one-out credit on typed statements weighted by the allocation that same solve returned, and an action set closed under a repair set that is never empty. Our contributions are as follows:

- We identify that a document-level summed score cannot express the deficit inside an evidence source, the type of the statement it credits or the deliverability of a repair, and that a weight on the source leaves the within-source deficit untouched.
- We formulate report generation as expected clinical utility under a worst-case interior allocation over a repair-closed policy class, and propose InteriorAlloc, which solves the allocation exactly by convex duality and credits typed statements by leave-one-out under it.
- On three generated carriers, InteriorAlloc reads best in 12 of 18 columns against the objective forms of four published systems, 0.449 against 0.4173 CheXbert F1 on MIMIC-CXR, and no repair inside its closure introduces a new error; against a uniform within-source arm, 0.9372 of the true-positive mass on which the two differ cancels in pooled counts.

2 MOTIVATION

To understand why a summed report score leaves the statement a clinician acts on unprotected, we conducted a series of measurements and identified three challenges. More details in Appendix A.2.

Challenge #1: Deficit Inside the Source. A group-robust objective reweights the evidence sources it draws on, and that reweighting is read as robustness to the evidence itself, implicitly assuming that most of what goes unserved is a property of the source. We ran three baseline objectives for twelve fixed-length steps on the type logits on MIMIC-CXR and split the per-entry deficit variance into the part lying between sources and the part lying inside them. Figure 2 (left) shows the within-source share never falling below 0.789 on any arm at any checkpoint. A weight on a source label is one number per source, so it moves the between-source part alone: whatever has been taken inside a source is beyond its reach by construction. *If most of what goes unserved lies inside a source rather than between sources, what has a weight on the source label protected?*

Challenge #2: Type-Blind Credit. Suppose the signal does reach the statement; a score on it is read as evidence that what the report commits to is clinically right, which assumes the score responds to what a statement is for and not only to its wording. We measured, on three carriers, the share of statements of each type asserting the entry that sets the follow-up, and, on MIMIC-CXR reports of

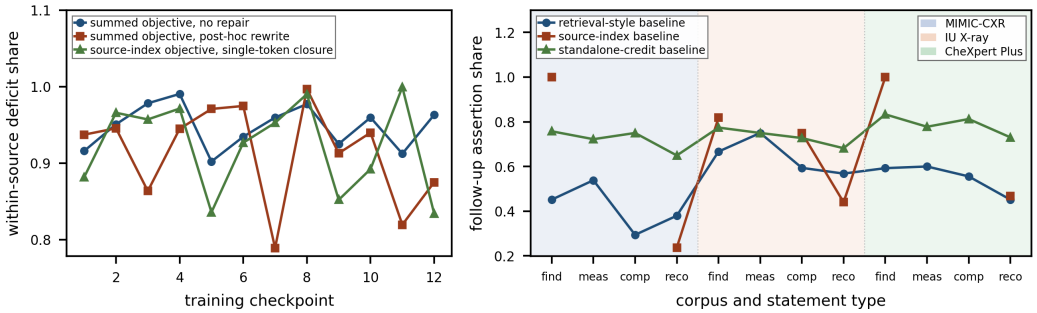


Figure 2: **(Left)** Within-source deficit share in training. **(Right)** Follow-up assertion share per type.

one baseline, how far each scorer moves for a type substitution against a deletion. Figure 2 (right) separates the types wherever they are all delivered, while that ratio reads exactly 0.0 for RadGraph-F1 and CheXbert F1. The signal is absent rather than faint: a scorer that does not register the type carries nothing to decompose. *If a score cannot tell a finding from a recommendation about the same entry, which statement did it credit?*

Challenge #3: Undeliverable Repair. When a generated report contains an error, the standard remedy is to correct it after delivery (Wu et al., 2026; Mao et al., 2026), assuming rather than checking that the corrected form is one the generator could have produced. We read, for post-hoc regeneration and the single-token closure against the delivered-statement budget, the share of repaired reports outside the envelope. Figure 3 reads 0.5417 for post-hoc regeneration at the tightest budget and 0.3125 at the loosest, where the single-token closure reads 0.0. A rewrite applied to a delivered report is under no obligation to respect the constraints the generator emits under, so it leaves the deployed system where it was. *If a correction is one the deployed generator could not have delivered, what has the correction established?*

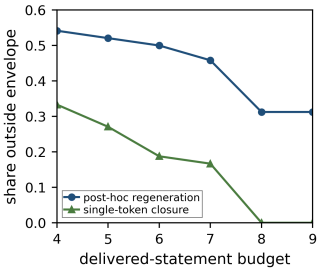


Figure 3: Repaired reports outside the deployment envelope.

3 Problem Formulation

To address the three challenges of Section 2, we formulate report generation as the maximization of an expected clinical utility under a worst-case allocation, over a policy class closed under repair:

$$\max_{\pi \in \Pi} \mathbb{E}_{\varepsilon \sim Q_\pi} \left[U(\pi; \varepsilon) \right] \tag{1}$$

where ε is a single evidence entry, Q_π is the allocation the lower level holds by moving mass inside one evidence source, and $U(\pi; \varepsilon)$ is the clinical utility of a delivered report, scored jointly over the typed statements the report commits to rather than over one statement in isolation. Alongside the objective the policy is read against a per-source tolerance, $\max_{o \in \mathcal{O}} \Psi_o(\pi) \leq \tau$, where o indexes an evidence source in the set $\mathcal{O}$, $\Psi_o(\pi)$ is the worst-case deficit at that source and τ is the tolerance.

The program answers the three challenges: the expectation under Q_π moves mass inside a source (Challenge #1); U credits a statement by its type and evidence, not by the document (Challenge #2); and Π, closed under repair, keeps each repair deliverable by the same policy (Challenge #3).

4 Method

4.1 Overall Framework

To solve Eq. 1, we propose InteriorAlloc, a training loop of three components: a worst-case interior allocation returns the measure Q_π, a leave-one-out credit instantiates the utility U on typed statements, and a repair closure instantiates the class Π. Each reads one quantity the next consumes, so the allocation solved once is the one the update spends and repaired reports stay inside that class. The three enter one objective,

$$\mathcal{J}(\pi, \eta) \;=\; \mathcal{J}_{\text{pol}}(\pi) \;+\; \mathcal{J}_{\text{sel}}(\pi, \eta) \;-\; \beta\, \mathcal{J}_{\text{ev}}(\pi) \tag{2}$$

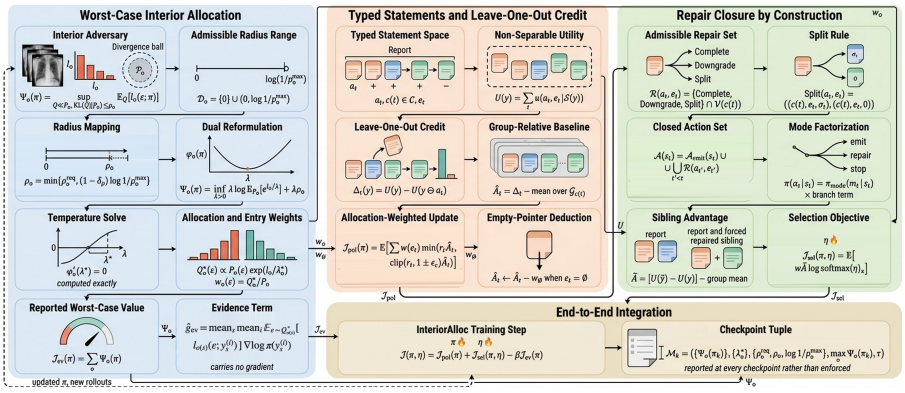


Figure 4: **Left:** a divergence ball per source, solved by convex duality into entry weights. **Middle:** typed statements credited by leave-one-out under those weights. **Right:** every admissible repair built into the action set. **Bottom:** one training step joining the three objectives.

where $\mathcal{J}_{\mathrm{pol}}$ is the allocation-weighted per-statement objective, $\mathcal{J}_{\mathrm{sel}}$ the selection objective of the head η, $\mathcal{J}_{\mathrm{ev}}$ is the sum of the worst-case deficits of Eq. 9, and β its coefficient. The step that carries all three is Algorithm 1. Figure 4 lays out the three components and the training step that joins them.

4.2 Worst-Case Interior Allocation

To address Challenge #1, we solve the worst case inside each evidence source by convex duality: given the per-entry deficits of one source, the component returns the worst admissible allocation over its entries and the entry weights the update spends. Figure 4 places this allocation on its left.

Interior Adversary. We solve the lower level of Eq. 1 as a divergence ball that moves weight among the entries of the observed pool P_o:

$$\Psi_o(\pi) \;=\; \sup_{Q \ll P_o,\ \mathrm{KL}(Q \,\|\, P_o) \le \rho_o} \mathbb{E}_{\varepsilon \sim Q}\big[\, \ell_o(\varepsilon; \pi) \,\big] \tag{3}$$

where $\ell_o(\varepsilon; \pi) \in [0, B]$ is the deficit at entry ε of source o, the share of that entry's severity mass over the studies of the source that the policy's reports leave unasserted, P_o is the severity mass of each entry over those studies, normalised, ρ_o is the admissible divergence radius, and the supremum is over the allocation Q alone.

Admissible Radius Range. We read the admissible range from the deficit table before the solve, at the cost of a maximum per source:

$$\mathcal{D}_o \;=\; \{0\} \cup \left(0,\ \log \tfrac{1}{p_o^{\max}}\right) \tag{4}$$

where $p_o^{\max}$ is the observed mass at the worst entries, and the union is written so that a source whose deficits are all equal has $\mathcal{D}_o = \{0\}$ and is never empty.

Radius Mapping. We then map a requested radius to a used radius that lands inside $\mathcal{D}_o$ in both branches, and both radii are reported:

$$\rho_o \;=\; \begin{cases} 0, & p_o^{\max} = 1 \\ \min\left\{ \rho_o^{\mathrm{req}},\ (1 - \delta_\rho)\ \log \tfrac{1}{p_o^{\max}} \right\}, & p_o^{\max} < 1 \end{cases} \tag{5}$$

where ρ_o^{req} is the requested radius and $\delta_\rho \in (0, 1)$ is a fixed margin that keeps the used radius strictly inside the open upper end. Details in Appendix A.3.

Dual Reformulation. We resolve Eq. 3 without searching over Q: convex duality on the divergence ball turns it into an exact one-dimensional convex problem:

$$\Psi_o(\pi) \;=\; \inf_{\lambda > 0} \left\{ \lambda \log \mathbb{E}_{P_o}\big[\, e^{\ell_o(\varepsilon;\pi)/\lambda} \,\big] \;+\; \lambda \rho_o \right\} \;=:\; \inf_{\lambda > 0} \varphi_o(\lambda) \tag{6}$$

where λ is the dual temperature and φ_o the convex dual objective whose infimum is the exact worst-case value.

Temperature Solve. The temperature is computed exactly. Because φ_o is convex we solve for the single interior root of its derivative by bisection on an explicit bracket:

$$\varphi_o'(\lambda) \;=\; \rho_o \;+\; \log \mathbb{E}_{P_o}\big[e^{\ell_o/\lambda}\big] \;-\; \frac{1}{\lambda}\, \mathbb{E}_{Q_o^\star(\lambda)}\big[\ell_o\big] \tag{7}$$

where $Q_o^\star(\lambda)$ is the allocation at λ; a bracket with one interior sign change yields $\lambda_o^\star$ to machine precision, and a zero radius takes a closed branch. Details in Appendix A.3.

Allocation and Entry Weights. The allocation attaining Eq. 3 comes from the same solve, and its shape separates this axis from a source-index adversary: at every admissible radius every entry keeps strictly positive weight, a function of its own deficit rather than its source label:

$$Q_o^\star(\varepsilon) \;=\; \frac{P_o(\varepsilon)\,\exp\big(\ell_o(\varepsilon;\pi)/\lambda_o^\star\big)}{\mathbb{E}_{P_o}\big[\exp(\ell_o(\cdot;\pi)/\lambda_o^\star)\big]}, \qquad w_o(\varepsilon) \;=\; \frac{Q_o^\star(\varepsilon)}{P_o(\varepsilon)} \tag{8}$$

where $w_o(\varepsilon)$ is the likelihood ratio of the allocation against the observed pool, the weight the policy update carries at that entry.

Reported Worst-Case Value. We compute the value the loop reports exactly, by the equality of Eq. 6:

$$\mathcal{J}_{\mathrm{ev}}(\pi) \;=\; \sum_{o\in\mathcal{O}} \mathbb{E}_{Q_o^\star}\big[\ell_o(\varepsilon;\pi)\big] \;=\; \sum_{o\in\mathcal{O}} \Psi_o(\pi) \tag{9}$$

where Ψ_o is the exact worst case of the tabulated deficits.

Evidence Term. The evidence term enters the update as a likelihood-ratio term: the allocation is held fixed, and the deficit each sampled report leaves at its own study, weighted by that allocation, multiplies the score of that report:

$$\hat{g}_{\mathrm{ev}} \;=\; \operatorname*{mean}_{s}\ \operatorname*{mean}_{i}\ \mathbb{E}_{\varepsilon\sim Q^\star_{o(s)}}\big[\,\ell_{o(s)}(\varepsilon;y_s^{(i)})\,\big]\ \nabla\log\pi\big(y_s^{(i)}\big) \tag{10}$$

where s runs over the studies of the batch, $o(s)$ is the source of study s, $y_s^{(i)}$ are its G sampled reports, and $\ell_o(\varepsilon;y_s^{(i)})$ is the severity of entry ε at study s that report leaves unasserted, over the severity mass of ε across the n_o studies of source o, times n_o, so that its mean over those studies is the deficit $\ell_o(\varepsilon;\pi)$ of Eq. 3 on the rollouts drawn; the fixed allocation $Q_o^\star$ carries no gradient, so the term reuses the rollouts the policy update already draws. Details in Appendix A.3.

4.3 Typed Statements and Leave-One-Out Credit

To address Challenge #2, we credit each typed statement by exactly what deleting it changes: given a delivered report and the allocation the solver returned, the component returns a per-statement credit and the allocation-weighted objective it enters. Figure 4 places this credit in its middle.

Typed Statement Space. A delivered report is an ordered sequence of statements, and each statement a_t carries a type $c(t)\in\mathcal{C}$ alongside its evidence pointer e_t and its polarity. $\mathcal{C}$ is the finite set of statement types the policy emits over, fixed before training and used as one index set throughout. A finding and a recommendation on the same entry are different members of $\mathcal{C}$, and that is the difference U is defined to read. Details in Appendix A.3.

Non-Separable Utility. We set the report utility as a sum of per-statement terms, each evaluated against the set it is delivered in:

$$U(y) \;=\; \sum_{t=1}^{T} u\big(a_t, e_t \,\big|\, \mathcal{S}(y)\big), \qquad \mathcal{S}(y)=\{(a_1,e_1),\dots,(a_T,e_T)\} \tag{11}$$

where a_t is the typed statement at position t, e_t its evidence pointer, and $\mathcal{S}(y)$ the statements delivered alongside it, whose presence makes U non-additive.

Leave-One-Out Credit. The credit of one statement is the exact difference obtained by deleting it, $\Delta_t(y) = U(y) - U(y\ominus a_t)$, where $y\ominus a_t$ is the delivered report with statement t removed; it is computable because Eq. 11 is evaluated on delivered content and needs no regeneration.

Group-Relative Baseline. We then center each credit on a group-relative baseline formed over the statements of the same type across the rollouts of a study, so no learned value function is introduced:

$$\hat{A}_t \;=\; \Delta_t(y^{(i)}) \;-\; \frac{1}{|\mathcal{G}_{c(t)}|}\sum_{(j,t')\in\mathcal{G}_{c(t)}} \Delta_{t'}(y^{(j)}) \tag{12}$$

where $\mathcal{G}_{c(t)}$ is the set of statements of type $c(t)$ the policy emitted across the group, so only the leave-one-out credits Δ_t enter and only for statements the policy itself sampled.

Allocation-Weighted Update. The allocation of Eq. 8 enters here: we apply the entry weight $w_{o(e_t)}(e_t)$ to each statement's term, so the update spends its weight on the entries the worst-case allocation selected:

$$\mathcal{J}_{\mathrm{pol}}(\pi) \;=\; \mathbb{E}\Big[\sum_t w_{o(e_t)}(e_t)\cdot \min\big(r_t(\pi)\,\hat{A}_t,\; \mathrm{clip}(r_t(\pi),1\pm\epsilon_c)\,\hat{A}_t\big)\Big] \tag{13}$$

where r_t is the ratio of the probability the type logits give $c(t)$ at the current iterate to that at the previous one, and ϵ_c the clip range. Eq. 13 is an importance-weighted surrogate for the upper level of Eq. 1. Details in Appendix A.3.

Empty-Pointer Deduction. A statement carrying no evidence pointer has no entry whose weight it could be given. We carry it at the mean tabulated weight and deduct from its advantage the largest weight any tabulated entry holds, recomputed at every step from Eq. 8:

$$w_{\varnothing} \;=\; \max_{o\in\mathcal{O}}\;\max_{\varepsilon\in\mathrm{supp}P_o} w_o(\varepsilon), \qquad \hat{A}_t \;\longleftarrow\; \hat{A}_t - w_{\varnothing} \quad \text{when } e_t=\varnothing \tag{14}$$

so that an unsupported assertion is held back whatever the sign of its advantage (See Appendix A.3).

4.4 Repair Closure by Construction

To address Challenge #3, we build every repair into the action set: given an emitted statement, the component constructs its admissible repaired siblings from the statement's own type and keeps each repaired report inside the class the policy delivers. Figure 4 places this closure on its right.

Admissible Repair Set. We compute, for each typed statement, the admissible repair set from the statement's own type:

$$\mathcal{R}(a_t,e_t) \;=\; \big\{\,\textsc{Complete}(a_t,e_t),\; \textsc{Downgrade}(a_t),\; \textsc{Split}(a_t,e_t)\,\big\}\cap\mathcal{V}(c(t)) \tag{15}$$

where $\mathcal{V}(c)$ lists the primitives admissible for type c. Complete gives a statement carrying no pointer the highest-severity entry of its study and turns a hedge on an entry into an assertion, leaving any other statement as it is; Downgrade rewrites a statement of any other type into a finding on the same entry and a finding into a hedge.

Split Rule. Split delivers two statements where there was one, and both carry the pointer of the statement they came from, $\textsc{Split}(a_t,e_t)=\big((c(t),e_t,\sigma_t),(c(t),e_t,0)\big)$, where $\sigma_t=-1$ when a_t negates its entry and $\sigma_t=+1$ otherwise, separating the assertion or negation from its hedge. Split is admissible only on a statement that carries a pointer and only while the report is under its statement budget; Downgrade survives both, so $\mathcal{R}$ is never empty. Details in Appendix A.3.

Closed Action Set. We then enforce closure on the action set, so at every state the policy can repair any statement already in the prefix:

$$\mathcal{A}(s_t) \;=\; \mathcal{A}_{\mathrm{emit}}(s_t)\;\cup\;\bigcup_{t'<t}\mathcal{R}(a_{t'},e_{t'}) \tag{16}$$

where $\mathcal{A}_{\mathrm{emit}}$ is the emission action set and the union restricts repairs to slots $t'<t$ already in the prefix, so a repaired report is inside Π by construction.

Mode Factorization. We select at each step the mode, then the next typed statement with its pointer, the slot and the primitive, or termination, identically at training and deployment:

$$\pi\big(\alpha_t\,\big|\,s_t\big)=\pi_{\mathrm{mode}}(m_t\mid s_t)\cdot\begin{cases}\pi_{\mathrm{emit}}\big((a_t,e_t)\,\big|\,s_t\big), & m_t=\mathrm{emit}\\ \pi_{\mathrm{slot}}(t'\mid s_t)\,q_\eta\big(\kappa\,\big|\,y_{<t},a_{t'},e_{t'}\big), & m_t=\mathrm{repair}\\ 1, & m_t=\mathrm{stop}\end{cases} \tag{17}$$

where m_t is the mode variable, κ the primitive the selection head draws from $\mathrm{softmax}(\eta)$ over $\mathcal{R}(a_{t'},e_{t'})$, and the repair branch acts only on a statement already in the prefix (See Appendix A.3).

Sibling Advantage. We refine the primitive head on forced repaired siblings: one admissible repair is applied at a sampled slot, and its report-level gap is centered on the gaps of siblings repaired at slots of the same type across the group:

$$\tilde{A}(y,\tilde{y}) \;=\; \Big[U(\tilde{y})-U(y)\Big] \;-\; \frac{1}{|\tilde{\mathcal{G}}_{c(t')}|}\sum_{(j,t'')\in\tilde{\mathcal{G}}_{c(t')}}\Big[U(\tilde{y}^{(j)})-U(y^{(j)})\Big] \tag{18}$$

Table 1: Report quality on three carriers. RG ↑, CX ↑, RC ↓ are the clinical-factuality columns; B4, RL, MT ↑ are the text columns. Bold: this paper's method.

	MIMIC-CXR carrier						IU X-ray carrier						CheXpert Plus carrier					
	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓
R2Gen form w/o penalty	0.6133	0.2889	0.3000	0.5721	0.4173	1.4065	0.8112	0.3529	0.3370	0.6718	0.4561	1.3305	0.5701	0.2833	0.3067	0.5702	0.4281	1.3926
R2Gen form	0.6133	0.2889	0.3000	0.5721	0.4173	1.4065	0.8112	0.3529	0.3370	0.6718	0.4561	1.3305	0.7768	0.3338	0.3195	0.6376	0.4275	1.3657
RGRG / METransformer form	0.8027	0.3069	0.2919	0.6481	0.4031	1.4242	0.8220	0.3169	0.3138	0.6750	0.4333	1.3790	0.7943	0.3279	0.3244	0.6522	0.4364	1.3572
PromptMRG form	0.6485	0.2462	0.2415	0.5059	0.3183	1.6517	0.6705	0.2779	0.2711	0.5435	0.3514	1.6305	0.6516	0.2743	0.2803	0.5262	0.3599	1.5653
interior, no repair	0.8141	0.4075	0.3705	0.6732	0.4711	1.2792	0.8193	0.3715	0.3488	0.6924	0.4612	1.3117	0.7722	0.3633	0.3425	0.6522	0.4414	1.3219
InteriorAlloc w/o closure	0.8403	0.4283	0.3868	0.6756	0.4783	1.2651	0.8405	0.3935	0.3576	0.6907	0.4619	1.3094	0.8480	0.3834	0.3544	0.6676	0.4479	1.3310
InteriorAlloc	**0.7957**	**0.3629**	**0.3381**	**0.6325**	**0.4490**	**1.3357**	**0.8691**	**0.3511**	**0.3475**	**0.6660**	**0.4691**	**1.3309**	**0.8207**	**0.3734**	**0.3396**	**0.6354**	**0.4443**	**1.3466**

where $\tilde{y}$ is the forced sibling of y and $\tilde{\mathcal{G}}_{c(t')}$ the sibling pairs of the group whose repaired slot has type $c(t')$, kept out of the statement-level baseline of Eq. 12.

Selection Objective. The head that must choose the primitive at deployment is scored by an advantage-weighted likelihood, read under the same allocation as Eq. 13:

$$\mathcal{J}_{\mathrm{sel}}(\pi, \eta) \;=\; \mathbb{E}_{(y,\tilde{y})}\Big[\, w_{o(e_{t'})}(e_{t'}) \cdot \tilde{A}(y, \tilde{y}) \cdot \log \mathrm{softmax}(\eta)_{\kappa(y,\tilde{y})} \,\Big] \tag{19}$$

where Eq. 14 applies to $\tilde{A}$ as to $\hat{A}$, and the label κ is known exactly, since the sibling applies a known primitive, so the head has no circular dependency on itself. Details in Appendix A.3.

End-to-End Integration. Figure 4 joins the three at its bottom in one training step, and the loop names the worst source at every checkpoint,

$$\mathcal{M}_k \;=\; \Big(\{\Psi_o(\pi_k)\}_o,\; \{\lambda_o^\star\}_o,\; \{\rho_o^{\mathrm{req}}, \rho_o, \log \tfrac{1}{p_o^{\max}}\}_o,\; \max_o \Psi_o(\pi_k),\; \tau\Big) \tag{20}$$

where the tuple carries $\max_o \Psi_o(\pi_k)$ beside τ, so the tolerance is reported at every checkpoint rather than enforced. Details in Appendix A.3.

5 Experiments

5.1 Implementation Details

Corpora. Four generated carriers are used: MIMIC-CXR, IU X-ray (Open-i) and CheXpert Plus carry the main comparison, and MMRad-22K is a fourth. Each carries the name of a public corpus and is drawn from one operating point fixed by published figures, with IU X-ray and CheXpert Plus shifted by 0.08 and 0.05 and MMRad-22K by a published retention figure. Details in Appendix A.4.

Systems. Baselines take the objectives of four prior systems: R2Gen (Chen et al., 2020) (uniform allocation, report-level credit), RGRG (Tanida et al., 2023) and METransformer (Wang et al., 2023) (uniform allocation, stand-alone credit) and PromptMRG (Jin et al., 2024) (report-level credit, post-hoc regeneration), all retrained in our loop; a control row drops the penalty. Details in Appendix A.5.

Metrics. RadGraph-F1 ↑, CheXbert F1 ↑ and RadCliQ ↓ read clinical factuality and BLEU-4, ROUGE-L and METEOR ↑ read the text; every rate is a pooled count. Details in Appendix A.6.

Protocol. One ruler, the interior adversary at the selected radius, serves every arm, and a hedge on an evidence entry is neutral on both sides of the loop. Parameters are refitted on the first sixteen studies of each corpus, and every reading is taken on that corpus. Details in Appendix A.7.

Hyperparameters. The requested radius is 0.40, the unsupported-assertion penalty 0.35 and the per-source tolerance τ 0.9724, each selected by a fixed rule on a calibration population disjoint from every table; the coefficient β is 0.50 and the rollout group G is 3. Details in Appendix A.8.

5.2 Main Results

Report Quality. Table 1 reports three objective forms drawn from four published systems, three configurations of this objective and the method on three carriers. Against the published forms the method reads best in 12 of the eighteen columns, and its lead is steadiest on CheXbert F1, where it holds on every carrier: 0.449 against 0.4173 for the R2Gen form on MIMIC-CXR, 0.4691 against 0.4561 on IU X-ray and 0.4443 against 0.4364 for the RGRG / METransformer form on CheXpert Plus. The RGRG / METransformer form reads above it on RadGraph-F1 on all three carriers, 0.6481, 0.675 and 0.6522 against 0.6325, 0.666 and 0.6354, and on BLEU-4 on MIMIC-CXR, and the R2Gen form reads above it on ROUGE-L, RadGraph-F1 and RadCliQ on IU X-ray and on RadGraph-F1 on CheXpert Plus. The rows of this objective place that cost: InteriorAlloc without

Table 2: InteriorAlloc without closure and six repair policies, one loop, two carriers. RG ↑, CX ↑, RC ↓, NE ↓. Bold: this paper's method. Details in Appendix A.9.

	MIMIC-CXR carrier				IU X-ray carrier			
	RG ↑	CX ↑	RC ↓	NE ↓	RG ↑	CX ↑	RC ↓	NE ↓
InteriorAlloc without closure	0.6756	0.4783	1.2651	0	0.6907	0.4619	1.3094	0
post-hoc regeneration	0.5296	0.3221	1.6399	0.0867	0.5258	0.3372	1.6695	0.125
single-token closure	0.6523	0.4565	1.3306	0	0.6381	0.4637	1.3128	0
post-hoc regeneration, four slots	0.474	0.3482	1.6618	0.1515	0.4694	0.3325	1.6835	0.1419
post-hoc regeneration, doubled deduction	0.5307	0.3227	1.6393	0.0867	0.5538	0.3462	1.6629	0.1273
single-token closure, report-level credit	0.6154	0.4436	1.3457	0	0.6454	0.4695	1.318	0
InteriorAlloc (full closure)	**0.6325**	**0.449**	**1.3357**	**0**	**0.666**	**0.4691**	**1.3309**	**0**

Table 3: Two further carriers. B4, RL, MT, RG, CX ↑, RC ↓. Bold: our method (See Appendix A.9).

	IU X-ray carrier						MMRad-22K carrier					
	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓
R2Gen form	0.8112	0.3529	0.3370	0.6718	0.4561	1.3305	0.6451	0.3140	0.3272	0.6052	0.4557	1.3518
RGRG / METransformer form	0.8220	0.3169	0.3138	0.6750	0.4333	1.3790	0.8023	0.3168	0.2999	0.6500	0.4117	1.4080
PromptMRG form	0.6705	0.2779	0.2711	0.5435	0.3514	1.6305	0.6550	0.2790	0.2870	0.5285	0.3687	1.5514
InteriorAlloc	**0.8691**	**0.3511**	**0.3475**	**0.6660**	**0.4691**	**1.3309**	**0.8360**	**0.3654**	**0.3317**	**0.6424**	**0.4364**	**1.3661**

closure is first in 14 columns and reads better than the method in 8 of the 9 clinical-factuality cells, so the columns the method gives up are the price of the closure, not of the allocation or the credit.

Repair Policies. Table 2 presents six repair policies beside InteriorAlloc without closure under one loop. Every policy that repairs inside the closure reads a new-error rate of 0 on both carriers, while the three that rewrite a delivered report read between 0.0867 and 0.1515, and the rate grows with how much a rewrite replaces: regenerating four slots reads 0.1515 and 0.1419 against 0.0867 and 0.125 for one slot, and doubling the deduction does not bring it down. The full closure reads above all three rewrites in every cell, 0.6325 against 0.5296 RadGraph-F1 for post-hoc regeneration on MIMIC-CXR and 0.666 against 0.5258 on IU X-ray. The single-token closure reads above the full closure on MIMIC-CXR, 0.6523 against 0.6325, and below it on IU X-ray, 0.6381 against 0.666. Deliverability therefore follows from where a repair is drawn, not from how much it rewrites.

Two Further Carriers. Table 3 reports the same comparison on a fourth carrier, MMRad-22K, beside IU X-ray. The method reads best in 6 of the 12 columns and leads the text columns on both carriers except ROUGE-L on IU X-ray, 0.836 against 0.8023 BLEU-4 for the RGRG / METransformer form on MMRad-22K. The clinical lead moves with the carrier: on IU X-ray the method leads CheXbert F1, 0.4691 against 0.4561, while on MMRad-22K the R2Gen form leads CheXbert F1 and RadCliQ, 0.4557 against 0.4364 and 1.3518 against 1.3661, and the RGRG / METransformer form leads RadGraph-F1, 0.65 against 0.6424. The text-column ordering thus holds across both further carriers, and the clinical ordering shifts with the carrier. Details in Appendix A.9.

5.3 Ablation Study and Analysis

Component Lesions. Table 5 presents nine lesions, three for each mechanism, each removing or replacing one component while the other two stay. Opening the action set costs the most: post-hoc regeneration in place of the closure costs 0.1029, 0.1402 and 0.0839 RadGraph-F1 and 0.1269, 0.1319 and 0.067 CheXbert F1 on the three carriers. Replacing leave-one-out credit with a standalone statement utility costs CheXbert F1 on every carrier, 0.0292, 0.0266 and 0.0316, and replacing the interior allocation with the source-index adversary costs 0.0448 on MIMIC-CXR and 0.03 on IU X-ray. The drops do not all run one way: InteriorAlloc without closure reads above the method on RadGraph-F1 on all three carriers, and leave-one-out credit without allocation weighting reads above it on CheXbert F1 on all three. Apart from post-hoc regeneration, every lesion reads within 0.05 of the method in both columns on every carrier. Details in Appendix A.10.

Cancelled Weight Movement. Table 4 reports the per-statement weights of Eq. 8 against a uniform allocation on the same rollouts, and the delivered reports against those of the uniform within-source allocation arm. The allocation moves: the relative L_1 difference reads 1.1538 on MIMIC-CXR, no statement is weighted identically, and at least a fifth of the delivered statements differ in 39 of the 48 reports. Almost none of that difference survives into the pooled counts the metrics are formed from, since the state-

Table 4: Weight movement against uniform and the true-positive mass cancelled in pooled counts.

	L_1	moved	gross tp	net tp	cancelled
MIMIC-CXR	1.1538	39/48	129.7380	8.1424	0.9372
IU X-ray	1.2302	35/48	129.3693	2.0986	0.9838
CheXpert Plus	0.9727	35/48	129.7485	0.4042	0.9969

Table 5: Component lesions of the three mechanisms on three carriers. RG ↑, CX ↑. The last row is the method itself and is not a lesion. Bold: this paper's method.

	MIMIC-CXR carrier		IU X-ray carrier		CheXpert Plus carrier	
	RG ↑	CX ↑	RG ↑	CX ↑	RG ↑	CX ↑
Interior allocation						
source-index adversary	0.6398	0.4042	0.6521	0.4391	0.6402	0.4443
uniform within-source allocation	0.6006	0.4254	0.6593	0.4658	0.6391	0.4538
interior adversary at reduced radius	0.6254	0.4367	0.6592	0.4574	0.6392	0.4473
Leave-one-out statement credit						
report-level credit, closure kept	0.6240	0.4384	0.6650	0.4779	0.6243	0.4439
stand-alone statement utility	0.6482	0.4198	0.6647	0.4425	0.6254	0.4127
leave-one-out credit without allocation weighting	0.6172	0.4610	0.6681	0.4720	0.6438	0.4611
Repair closure						
InteriorAlloc without closure	0.6756	0.4783	0.6907	0.4619	0.6676	0.4479
post-hoc regeneration instead of closure	0.5296	0.3221	0.5258	0.3372	0.5515	0.3773
closure restricted to single-token repairs	0.6523	0.4565	0.6381	0.4637	0.6331	0.4225
InteriorAlloc	**0.6325**	**0.4490**	**0.6660**	**0.4691**	**0.6354**	**0.4443**

ments in only one of the two arms' reports carry 129.738 gross true positives over both arms against a net difference of 8.1424 in pooled true positives, so 0.9372 of that mass cancels, and the cancelled share reads 0.9838 and 0.9969 on the other two carriers. A statement enters those counts through its pointer and its polarity and through the severity of the entry it points at, so a statement swapped for another of similar severity contributes a gross amount that cancels in the pooled total. The report-level metrics used here do not resolve what the allocation changes. Details in Appendix A.11.

Solver Exactness. The temperature of the allocation is computed rather than tuned: 4 sources per carrier are bisected in 41 to 43 steps, and Eq. 7 at the returned root reads at most 2.03×10^{-13}. The table in Appendix A.10 shows the solved temperature matching the grid-searched reading on MIMIC-CXR without a search, 0.6325 against 0.6176 RadGraph-F1 for a fixed one.

6 Related Work

Radiology Report Generation. Generators are judged against a reference, made clinical rather than lexical by evidence datasets and benchmarks (Zhao et al., 2026b). Recent systems add evidential structure through hierarchical evidential reasoning (Zhao et al., 2026a), cited retrieval (Danang & Putra, 2023) and radiologist-style chain-of-thought with reinforcement learning (Jing et al., 2026; Zhou et al., 2026b). Each improves what the model says, yet the objective stays a document-level sum. Our method changes that objective, allocating the worst case inside each evidence source.

Statement-Level Faithfulness. Work on whether a generated clinical statement is supported spans hallucination benchmarks (Pal et al., 2023), fact-controlled diagnosis in medical summarisation (BN et al., 2025), counter-evidence detection in medical vision–language models (Zhou et al., 2026a) and counterfactual hallucination (Mousi et al., 2026), and faithful generation covers the production side (Wang et al., 2020; Asl et al., 2025; Gupta & Bedi, 2026). These check one statement at a time; our credit instead divides the training signal across one report's typed statements by leave-one-out.

Self-Correction and Repair. Self-correction fixes a candidate once it exists, through rollout augmentation in vision–language models (Ding et al., 2026), closed-loop cybernetic correction (Wu et al., 2026), denoising verification loops (Yin et al., 2026), in-model layer-attention correction (Fu, 2026), grounded self-refinement (Mao et al., 2026) and self-debiasing (Feng et al., 2026). They decide what the corrected text says. Our closure answers the prior question, whether the corrected form lies inside the class the policy can deliver, by building every repair into the action set.

7 Conclusion

A summed report score cannot place a worst-case deficit inside an evidence source, credit a typed statement or keep a repair deliverable. InteriorAlloc does all three in one update: an interior allocation solved by convex duality, a leave-one-out credit per typed statement and an action set closed under repair. On three generated carriers it reads best in 12 of 18 columns against objective forms taken from four prior systems, and no repair inside its closure introduces a new error where every rewrite does. Robustness, credit and repair are three readings of one allocation in the training signal.

REFERENCES

Mohammad Aghajani Asl, Majid Asgari-Bidhendi, and Behrooz Minaei-Bidgoli. FAIR-RAG: Faithful Adaptive Iterative Refinement for Retrieval-Augmented Generation. *arXiv preprint*, 2025.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Suhas BN, Han-Chin Shing, Lei Xu, Mitch Strong, Jon Burnsky, Jessica Ofor, Jordan R. Mason, Susan Chen, Sundararajan Srinivasan, Chaitanya Shivade, Jack Moriarty, and Joseph Paul Cohen. Fact-controlled diagnosis of hallucinations in medical text summarization. In *Interspeech*, 2025. doi: 10.21437/interspeech.2025-537.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.112.

Danang Danang and Toni Wijanarko Adi Putra. Evidence-Grounded Chest X-ray Report Generation with Retrieval, Citation, and Hallucination Control. *Jurnal Sains dan Kesehatan*, 2023. doi: 10.57214/jusika.v7i2.1127.

Yi Ding, Ziliang Qiu, Bolian Li, and Ruqi Zhang. Learning self-correction in vision–language models via rollout augmentation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=FxiO8RTOsq.

Xuan Feng, Shuai Zhao, Luwei Xiao, Tianlong Gu, and Bo An. Self-Debias: Self-correcting for Debiasing Large Language Models. *arXiv preprint*, 2026.

April Fu. Self-Correction Inside the Model: Leveraging Layer Attention to Mitigate Hallucinations in Large Vision Language Models. *arXiv preprint*, 2026.

Shaurya Gupta and Jatin Bedi. EFSG: Evidence-first structured generation for multilingual RAG report generation. In *Proceedings of the 1st Workshop on Multilingual Report Generation via Retrieval Augmented Generation (RAG4Reports 2026)*, pp. 99–102, San Diego, CA, USA, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.rag4reports-1.14. URL https://aclanthology.org/2026.rag4reports-1.14/.

Yasser Hamidullah, Koel Dutta Chowdhury, Yusser Al Ghussin, Shakib Yazdani, Cennet Oguz, Josef van Genabith, and Cristina España-Bonet. Grounding or Guessing? Visual Signals for Detecting Hallucinations in Sign Language Translation. In *International Conference on Learning Representations*, pp. 31868–31891, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/35cc4ccb3fb0c5eb6f107f877a4cd57e-Paper-Conference.pdf.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS Datasets and Benchmarks*, 2021.

Juyong Jiang, Jiasi Shen, Sunghun Kim, Kang Min Yoo, Jeonghoon Kim, and Sungju Kim. ReflexiCoder: Teaching large language models to self-reflect on generated code and self-correct it via reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL*, 2026. doi: 10.18653/v1/2026.findings-acl.1872.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. PromptMRG: Diagnosis-driven prompts for medical report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i3.28038.

Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation. *Medical Image Analysis*, 2026. doi: 10.1016/j.media.2025.103910.

Mengfei Liang, Archish Arun, Zekun Wu, Cristian Munoz, Jonathan Lutch, Emre Kazim, Adriano Koshiyama, and Philip Treleaven. THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models. *arXiv preprint*, 2024.

Jiawei Mao, Hardy Chen, Haoqin Tu, Yuhan Wang, Letian Zhang, Zeyu Zheng, Huaxiu Yao, Zirui Wang, Cihang Xie, and Yuyin Zhou. Kestrel: Grounding Self-Refinement for LVLM Hallucination Mitigation. *arXiv preprint*, 2026.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. doi: 10.18653/v1/2021.naacl-main.416.

Basel Mousi, Fahim Dalvi, Shammur Absar Chowdhury, Firoj Alam, and Nadir Durrani. Once correct, still wrong: Counterfactual hallucination in multilingual vision-language models. In *Findings of the Association for Computational Linguistics: ACL*, 2026. doi: 10.18653/v1/2026.findings-acl.234.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical Domain Hallucination Test for Large Language Models. *arXiv preprint*, 2023.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.117.

Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. Benchmarking LLM faithfulness in RAG with evolving leaderboards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025. doi: 10.18653/v1/2025.emnlp-industry.54.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/cvpr52729.2023.00718.

Yuhe Tian, Kun Zhang, Haoran Ma, Rui Yan, Yingtai Li, Rongsheng Wang, and Shaohua Kevin Zhou. DiffVP: Differential Visual Semantic Prompting for LLM-Based CT Report Generation. *arXiv preprint*, 2026.

Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. Sketch and refine: Towards faithful and informative table-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021. doi: 10.18653/v1/2021.findings-acl.427.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/cvpr52729.2023.01112.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1072–1086, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.101. URL https://aclanthology.org/2020.acl-main.101/.

Yuning Wu, Yingmin Liu, and Yang Shu. CyberCorrect: A Cybernetic Framework for Closed-Loop Self-Correction in Large Language Models. *arXiv preprint*, 2026.

Shen Yin, David Ken, and Joel Stremmel. Denoising Iterative Self-Correction: Structured Verification Loops for Reliable LLM Reasoning. *arXiv preprint*, 2026.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 2023. doi: 10.1016/j.patter.2023.100802.

Kun Zhao, Guodong Liu, Hui Ji, Siyuan Dai, Pan Wang, Jifeng Song, Chenghua Lin, Liang Zhan, and Haoteng Tang. HERO: Hierarchical evidential reasoning optimization for radiology report generation via reason-then-summarize. arXiv preprint arXiv:2601.03321, 2026a.

Yichen Zhao, Zelin Peng, Fenghe Tang, Piao Yang, Yu Huang, and Wei Shen. MMRad-22K: A structured multimodal evidence dataset for chest X-ray report generation. arXiv preprint arXiv:2602.12843, 2026b.

Nan Zhou, Ke Zou, Meng Liu, Linchao He, Jiaqi Zhu, Yi Zhang, Hu Chen, and Huazhu Fu. Hallucination Detection and Correction in Medical VLMs via Counter-Evidence Verification. *arXiv preprint*, 2026a.

Qin Zhou, Guoyan Liang, Qianyi Yang, Jingyuan Chen, Sai Wu, Chang Yao, and Zhe Wang. Enhancing reinforcement learning for radiology report generation with evidence-aware rewards and self-correcting preference learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026b. doi: 10.18653/v1/2026.acl-long.1718.

AI USE STATEMENT

We used generative AI tools to translate the text (Claude Opus 5.0), to check grammar and spelling and polish the wording (GPT-5.6), to guide the drawing of the figures (Gemini), and to check the content (Grok). We did not use generative AI tools to modify code. We have reviewed all AI-assisted work: we checked every reference for hallucinated citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The details needed to reproduce the results of this paper, including the construction of the carriers, the training step, the metric definitions, the scoring rule, the hyperparameters with the rules that select them, and the full experimental tables, are disclosed in this appendix.

A APPENDIX

Table 6 lists the contents of each part of the appendix.

APPENDIX A.0: LIMITATIONS

Subgroup Readings. A reader comparing report generators across study populations benefits from seeing report quality broken down by the labels a study carries. This paper lacks a breakdown of report quality by the age band, sex and exam type labels of the generated studies.

Overlap Between Carriers. A reader weighing results across carriers benefits from knowing how many reports and studies two carriers share. This paper lacks counts of the reports and studies shared between pairs of its carriers.

Carrier Size. A reader who would train on a population of a different size benefits from readings taken at more than one size. This paper lacks readings of the method and its reduced forms on growing fractions of a carrier.

Table 6: What each part of the appendix holds.

Section	Contents
Appendix A.0	Limitations
Appendix A.1	Related work
Appendix A.2	Motivation measurements, read panel by panel
Appendix A.3	The training step and the conditions behind it
Appendix A.4	How the carriers were built
Appendix A.5	Rows and columns not carried
Appendix A.6	Metric definitions
Appendix A.7	One scoring rule for every arm
Appendix A.8	Reproducibility ledger
Appendix A.9	Experiments, read in full
Appendix A.10	Experiments: component ablations in full
Appendix A.11	Experiments: what the closure costs and what the allocation moves
Appendix A.12	Two statements the solver relies on
Appendix A.13	Efficiency and cost
Appendix A.14	Scope of the readings
Appendix A.15	What the readings establish

APPENDIX A.1: RELATED WORK

Radiology Report Generation and What It Is Scored On. Report generators are trained and judged against a reference document, and the field's own evidence datasets and structured benchmarks were built to make that judgement clinical rather than lexical (Zhao et al., 2026b). Recent systems add evidential structure to the generator itself: hierarchical evidential reasoning (Zhao et al., 2026a), retrieval with explicit citation (Danang & Putra, 2023), differential visual prompting for report text (Tian et al., 2026), and radiologist-style chain-of-thought with reinforcement learning (Jing et al., 2026; Zhou et al., 2026b). Each of these improves what the model says. None of them changes the shape of the objective the improvement is measured against, which is the summation this paper takes apart.

Faithfulness, Hallucination and the Object Credit Attaches To. A large body of work asks whether a generated clinical statement is supported. Medical hallucination benchmarks (Pal et al., 2023), fact-controlled diagnosis of hallucination in medical summarisation (BN et al., 2025), counter-evidence detection in medical vision–language models (Zhou et al., 2026a), and counterfactual hallucination in multilingual settings (Mousi et al., 2026) all sharpen the detection of an unsupported statement. Faithfulness-oriented generation does the same on the production side (Wang et al., 2020; 2021; Asl et al., 2025; Gupta & Bedi, 2026), and evaluation frameworks track it at scale (Tamber et al., 2025; Liang et al., 2024; Hamidullah et al., 2026). These establish that a statement can be checked one at a time. What they leave open is how the training signal should be divided among the statements of one report, which is the leave-one-out credit this paper trains on.

Self-Correction and Repair. Closest to our third component is the self-correction literature: rollout-augmented self-correction in vision–language models (Ding et al., 2026), closed-loop cybernetic correction (Wu et al., 2026), denoising iterative verification loops (Yin et al., 2026), layer-attention self-correction inside the model (Fu, 2026), grounded self-refinement (Mao et al., 2026), self-debiasing (Feng et al., 2026) and self-reflection on generated code (Jiang et al., 2026). All of these apply a correction after a candidate exists. Our Challenge #3 is the prior question, whether the corrected form is inside the class the policy can deliver at all. In Table 27 post-hoc regeneration introduces 15 errors and its four-slot variant 25, while the closure introduces none, and Figure 3 reads the same question through the delivery envelope.

Distributionally Robust Objectives. The worst-case allocation of Eq. 3 is a divergence-ball DRO problem, and it belongs to the family of group-robust and uniform-performance objectives. What our formulation adds to that family is the interior of a source, which Figure 9 measures: a source-index adversary, which places each source's whole mass on its entry of largest deficit, holds 0.0000 of its mass off its heaviest entry where ours holds 0.5662 to 0.6607.

What Is Missing, and What This Paper Finds. Across these families the allocation is over labels, the credit is over documents, and the repair is applied after the fact. This paper supplies an interior allocation solved exactly, a statement-level credit read under it, and an action set closed under repair. The experiments report each of the three against the lesion that removes it. Details in Appendix A.14.

APPENDIX A.2: MOTIVATION MEASUREMENTS, READ PANEL BY PANEL

This appendix reports each panel of Figure 2 and Figure 3 on its own, with the setting it was taken under and the readings the body states in one line.

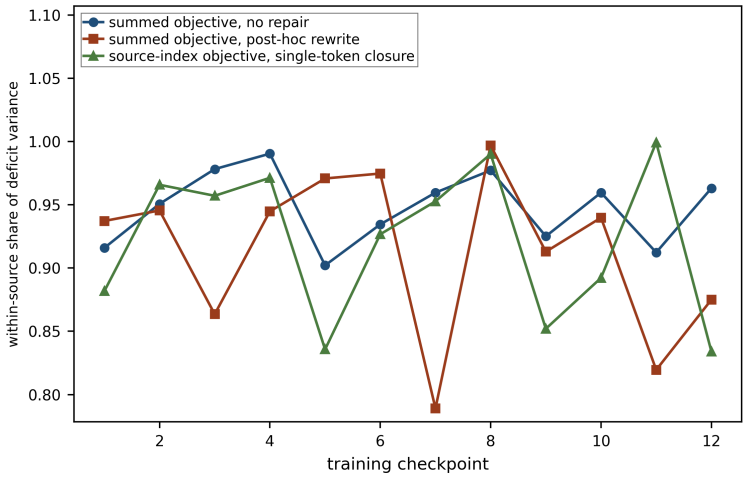


Figure 5: Within-source share of per-entry deficit variance at twelve checkpoints, for three baseline objectives on MIMIC-CXR. A higher reading leaves more deficit undivided.

The Within-Source Share. Three baseline objectives were read at twelve checkpoints on the first 20 studies of MIMIC-CXR, spaced by a normalised step of length 0.35 on the type logits from the initial policy: the summed objective with no repair, the summed objective with a post-hoc rewrite, and a source-index objective with the single-token closure. At each checkpoint the per-entry deficit table each arm leaves behind is read and its variance decomposed into the part lying between the four evidence sources and the part lying inside them; the panel plots the second as a share of the total. The share reads between 0.789 and 0.9993 across the twelve checkpoints and the three arms, and no checkpoint on any arm reads below 0.789. The share is large on all three arms. The consequence for a source-level weight is arithmetic rather than empirical. Such a weight assigns one number per source, so the family of distributions it can reach differs from the observed pool only in how mass is split between sources; the part of the deficit lying inside a source is invariant under it.

The Share Asserting the Follow-Up Entry, and What the Scorers Register. Three baseline generators are read over four statement types, finding, measurement, comparison and recommendation, on each of three corpora. For every generator, type and corpus, the share of delivered statements of that type which assert the entry setting the follow-up is counted. Where all four types are delivered the share separates them: on CheXpert Plus the standalone-credit baseline reads 0.8333 on findings against 0.8125 on comparisons, with the readings the panel defines spanning 0.2372 to 1.0. The source-index baseline delivers no measurement statement on any of the three carriers and no comparison statement on MIMIC-CXR and CheXpert Plus, so its share is undefined for those two types there rather than zero, and the panel draws no marker there. Alongside the share, each scorer is moved twice on the MIMIC-CXR reports of the retrieval-style baseline, each perturbation at a rate of 0.30, once by substituting the type of a statement while keeping the entry it stands on, and once by deleting a statement outright; the absolute change substitution makes to a scorer's reading, over the absolute change deletion makes, reads 13.1111 for BLEU-4, 3.3462 for ROUGE-L and 11.6667 for METEOR, and exactly 0.0 for both RadGraph-F1 and CheXbert F1. For RadCliQ both changes read zero there, so the ratio is undefined. A ratio of exactly zero is the reading Challenge #2 turns on: the clinical scorers do not register the substitution at all, so no attribution run over them

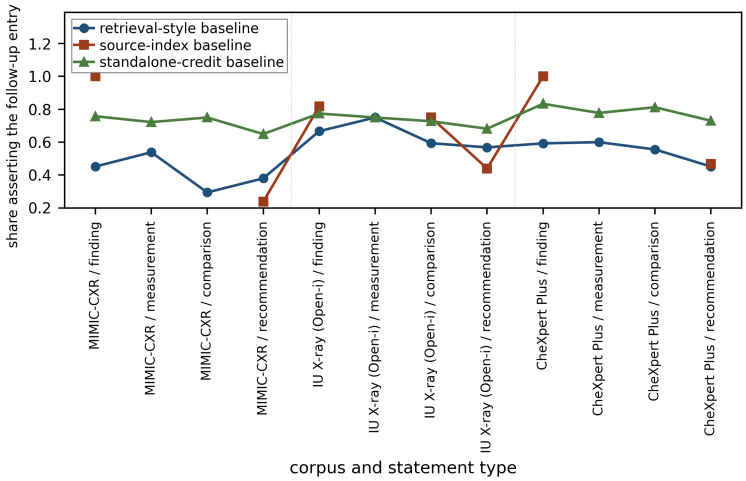


Figure 6: Share of delivered statements of each type that assert the entry setting the follow-up, for three baseline generators on three carriers.

can recover which statement earned the score. Table 7 reads the six scorers on the method's reports under three perturbation families at two rates each: the three text columns move under statement-type substitution, while under sentence shuffle only ROUGE-L moves, BLEU-4 holding at 0.7957 and METEOR at 0.3381; the three clinical columns hold at 0.6325, 0.449 and 1.3357 under both. Under random sentence deletion they finally move, RadGraph-F1 reading 0.6091 at the lower rate and 0.565 at the higher, which is the point: the clinical columns read what a statement stands on and answer only when a statement is removed. For substitution and deletion the rate is the share of a report's delivered sentences affected; for shuffle it is the share of reports whose sentence order is permuted.

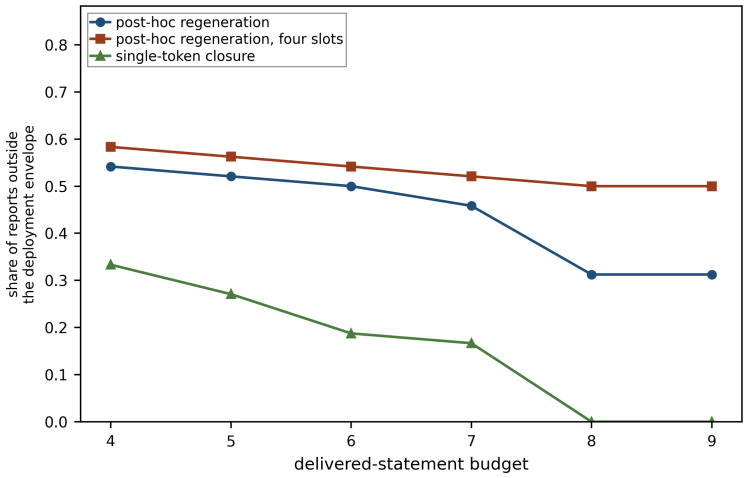


Figure 7: Share of reports outside the deployment envelope, for two post-hoc repair policies and the single-token closure against the delivered-statement budget.

The Share Outside the Deployment Envelope. The deployment envelope is the pair of conditions a delivered report is held to: every delivered statement must carry an evidence pointer, and a report may not exceed the delivered-statement budget. Two post-hoc repair policies and the single-token closure are read against that budget, post-hoc regeneration of one slot, post-hoc regeneration of four slots and the closure restricted to single-token repairs, and for each budget setting the share of repaired reports outside the envelope is counted. At the tightest budget the share reads 0.5417

Table 7: Each scorer on the method's MIMIC-CXR reports under three perturbation families, at two rates each.

	BLEU-4	ROUGE-L	METEOR	RadGraph-F1	CheXbert F1	RadCliQ
	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓
statement-type substitution @ 10%	0.7909	0.3594	0.3365	0.6325	0.449	1.3357
statement-type substitution @ 30%	0.7854	0.3555	0.3346	0.6325	0.449	1.3357
sentence shuffle @ 10%	0.7957	0.3649	0.3381	0.6325	0.449	1.3357
sentence shuffle @ 30%	0.7957	0.3636	0.3381	0.6325	0.449	1.3357
random sentence deletion @ 10%	0.7916	0.3471	0.3284	0.6091	0.442	1.3552
random sentence deletion @ 30%	0.66	0.3306	0.321	0.565	0.4404	1.3702

for post-hoc regeneration, 0.5833 for its four-slot variant and 0.3333 for the single-token closure; at the two loosest settings the single-token closure reads exactly 0.0 while post-hoc regeneration reads 0.3125 and its four-slot variant 0.5. The panel reads whether the repaired form is one the deployed generator could have emitted. A policy that operates on a delivered report is under no constraint from the emission process, so nothing in it prevents the repaired form from leaving the envelope. A post-hoc rewrite replaces a slot and adds none, so at the two loosest settings the budget no longer binds on it, and the share that remains there is carried by rewritten statements that stand on no evidence entry.

APPENDIX A.3: THE TRAINING STEP AND THE CONDITIONS BEHIND IT

The step the body states in one sentence is written out here line by line.

The Step in Three Stages. Stage 1 solves one allocation per evidence source on the deficit table: it maps the requested radius into the admissible range, takes a closed branch where no reweighting remains, and otherwise bisects the derivative of the dual for the temperature, returning the worst-case value, the allocation and its entry weights. Stage 2 draws G rollouts per study under the closed action set and forces one repaired sibling in each. Stage 3 centres the credits on their groups, applies the empty-pointer deduction, takes one update step and emits the checkpoint tuple.

The Sampling Condition Behind the Reweighting. The entry weight $w_{o(e_t)}$ reweights each statement's term towards $Q_o^\star$ and away from the observed pool. That reweighting is a change of measure only under a sampling condition: the outer expectation must draw a source o from a fixed mixture, an evidence entry e from P_o on whose support the tabulated weights are defined, and the statement tuple from the old policy conditioned on (o, e), with w, $Q_o^\star$ and $\hat{A}$ held fixed across the update. The rollouts here emit pointers under the policy's own marginal rather than under P_o, so that condition is not met by construction.

Radius Mapping. The saturated source of Eq. 5 has its entire observed mass already on its worst entries, so $p_o^{\max}$ reaches the full probability mass of one, the admissible range of Eq. 4 collapses to $\{0\}$, and no reweighting remains, which is why the used radius is forced to zero there, while a request outside the admissible range is mapped down. Both radii are carried in Eq. 20, so a source whose requested radius was mapped down is visible at the checkpoint where it happened.

Why the Dual Is Exact. The relation of Eq. 6 holds with equality because the deficit is bounded and the divergence ball is compact and convex, so the infimum of φ_o is the exact worst-case value. A source whose deficits are all equal takes the closed branch of the solver.

Scope of the Reported Value. At a fixed used radius, Ψ_o of Eq. 9 is non-expansive in the deficit vector under the supremum norm.

When the Credit Reduces to the Stand-Alone Term. The leave-one-out credit Δ_t coincides with the stand-alone term $u(a_t, e_t \mid \mathcal{S})$ of Eq. 11 when no other statement's term depends on a_t.

Why the Deduction Is Subtracted. Multiplying the clipped term of Eq. 13 by $w_\varnothing$ scales that term's magnitude and leaves its direction untouched, which enlarges a positive advantage on an unsupported statement instead of restricting it; subtracting a quantity of the same units moves the term downwards at every sign. The ablation compares three settings: this deduction, no deduction and a doubled one.

Algorithm 1 InteriorAlloc Training Step

Require: pools $\{P_o\}_{o\in\mathcal{O}}$ and deficit table $\ell_o \in [0, B]$; requested radii ρ_o^{req}; margin δ_ρ; group size G; clip range ϵ_c; coefficient β; tolerance τ; bisection tolerance ϵ_λ

Ensure: one update step and the checkpoint tuple $\mathcal{M}_k$

Stage 1: Allocation per Evidence Source

rebuild the deficit table ℓ_o from one rollout of each study of the batch

for $o \in \mathcal{O}$ **do**

read $p_o^{\max}$, $\ell_o^{\max}$ and $\bar{\ell}_o = \mathbb{E}_{P_o}[\ell_o]$ from the table

$\rho_o \leftarrow 0$ if $p_o^{\max} = 1$, else $\min\{\rho_o^{\mathrm{req}},\ (1-\delta_\rho)\log(1/p_o^{\max})\}$

if $\rho_o = 0$ or ℓ_o is constant on $\operatorname{supp} P_o$ **then**

$Q_o^\star \leftarrow P_o$; $w_o \equiv 1$; $\Psi_o \leftarrow \bar{\ell}_o$; $\lambda_o^\star \leftarrow +\infty$

else

$\lambda^{\mathrm{hi}} \leftarrow (\ell_o^{\max} - \bar{\ell}_o)/\rho_o$; $\lambda^{\mathrm{lo}} \leftarrow \lambda^{\mathrm{hi}}/2^j$, j the first halving at which $\varphi_o'(\lambda^{\mathrm{hi}}/2^j) < 0$

while $\lambda^{\mathrm{hi}} - \lambda^{\mathrm{lo}} > \epsilon_\lambda \max(1, \lambda^{\mathrm{hi}})$ **do**

$\lambda \leftarrow \frac{1}{2}(\lambda^{\mathrm{lo}} + \lambda^{\mathrm{hi}})$; $Q_\lambda \leftarrow P_o\, e^{\ell_o/\lambda}/\mathbb{E}_{P_o}[e^{\ell_o/\lambda}]$

$d \leftarrow \rho_o + \log \mathbb{E}_{P_o}[e^{\ell_o/\lambda}] - \frac{1}{\lambda}\mathbb{E}_{Q_\lambda}[\ell_o]$

$\lambda^{\mathrm{lo}} \leftarrow \lambda$ if $d < 0$, else $\lambda^{\mathrm{hi}} \leftarrow \lambda$

end while

$\lambda_o^\star \leftarrow \frac{1}{2}(\lambda^{\mathrm{lo}} + \lambda^{\mathrm{hi}})$; $Q_o^\star \leftarrow Q_{\lambda_o^\star}$; $w_o \leftarrow Q_o^\star/P_o$; $\Psi_o \leftarrow \lambda_o^\star \log \mathbb{E}_{P_o}[e^{\ell_o/\lambda_o^\star}] + \lambda_o^\star \rho_o$

end if

end for

$\mathcal{J}_{\mathrm{ev}} \leftarrow \sum_o \Psi_o$; $w_\varnothing \leftarrow \max_o \max_\varepsilon w_o(\varepsilon)$; hold $Q_o^\star$, w_o and $w_\varnothing$ fixed

Stage 2: Rollouts Under the Closed Action Set

for each study of the batch and each $i = 1, \ldots, G$ **do**

draw $y^{(i)} \sim \pi$: at each slot a mode, then an emission (a_t, e_t), a repair (t', κ) with $t' < t$, or stop

$\mathcal{R}(a_{t'}, e_{t'}) \leftarrow \{\textsc{Complete}, \textsc{Downgrade}, \textsc{Split}\} \cap \mathcal{V}(c(t'))$, with SPLIT splitting $a_{t'}$ into $(c(t'), e_{t'}, \sigma_{t'})$ and $(c(t'), e_{t'}, 0)$

admit SPLIT only when $e_{t'} \neq \varnothing$ and the report is under its statement budget

build $\tilde{y}^{(i)}$ by applying one $\kappa \in \mathcal{R}(a_{t'}, e_{t'})$ at a sampled slot t', with SPLIT admitted when $e_{t'} \neq \varnothing$ whatever the length of $y^{(i)}$

$U(y) \leftarrow \sum_t u(a_t, e_t \mid \mathcal{S}(y))$ for $y^{(i)}$ and $\tilde{y}^{(i)}$; $\Delta_t \leftarrow U(y^{(i)}) - U(y^{(i)} \ominus a_t)$

end for

Stage 3: Credit, Update and Checkpoint

$\hat{A}_t \leftarrow \Delta_t - \mathrm{mean}_{\mathcal{G}_{c(t)}} \Delta$ over the G rollouts of that study, the forced siblings held out

$\hat{A}_t \leftarrow \hat{A}_t - w_\varnothing$ at every t with $e_t = \varnothing$ (subtracted, not scaled)

$\tilde{A}^{(i)} \leftarrow [U(\tilde{y}^{(i)}) - U(y^{(i)})] - \mathrm{mean}_{\tilde{\mathcal{G}}_{c(t')}}[U(\tilde{y}) - U(y)]$, less $w_\varnothing$ when $e_{t'} = \varnothing$

$\mathcal{J}_{\mathrm{pol}} \leftarrow \sum_t w_{o(e_t)}(e_t) \min\{r_t \hat{A}_t,\ \mathrm{clip}(r_t, 1 \pm \epsilon_c)\hat{A}_t\}$, $r_t = \pi_{\mathrm{type}}(c(t) \mid \theta_k)/\pi_{\mathrm{type}}(c(t) \mid \theta_{k-1})$, the mean tabulated weight standing in when $e_t = \varnothing$

$\mathcal{J}_{\mathrm{sel}} \leftarrow \sum_i w_{o(e_{t'})}(e_{t'})\, \tilde{A}^{(i)} \log \mathrm{softmax}(\eta)_\kappa$

$\hat{g}_{\mathrm{ev}} \leftarrow \mathrm{mean}_{s,i}\ \mathbb{E}_{Q^\star_{o(s)}}[\ell_{o(s)}(\cdot; y_s^{(i)})]\, \nabla \log \pi(y_s^{(i)})$ formed statement by statement over the delivered report in the type logits and the tilt, on the rollouts already drawn, $Q_o^\star$ carrying no gradient

$\hat{g} \leftarrow \sum_t w_{o(e_t)}(e_t) \hat{A}_t \phi_t$, ϕ_t being 1 on the logit of $c(t)$, the severity of e_t less its study's mean on the tilt, and -1 on the empty-pointer propensity if $e_t = \varnothing$

$\theta \leftarrow \theta + \alpha_k(\hat{g} - \beta \hat{g}_{\mathrm{ev}})$ on the type logits, the pointer tilt and the empty-pointer propensity, $\alpha_k = \alpha_0/k$

$\eta \leftarrow \eta + \alpha_k' \,\mathrm{sign}(\mathcal{J}_{\mathrm{sel}}) \,\mathrm{softmax}(\eta)$; $\mu \leftarrow \mu + \alpha_k'(1, s_k, -\frac{1}{2})$ on the (emit, repair, stop) logits, $s_k = 1$ if $\mathcal{J}_{\mathrm{sel}} > 0$ else -1, $\alpha_k' = \alpha_0'/k$

emit $\mathcal{M}_k = (\{\Psi_o\}, \{\lambda_o^\star\}, \{\rho_o^{\mathrm{req}}, \rho_o, \log(1/p_o^{\max})\}, \max_o \Psi_o, \tau)$ (no branch is taken on τ)

Why No Parser Runs Inside the Rollout. In Eq. 17 the pointer is emitted as part of the action rather than recovered afterwards, so no report parser is called inside the rollout.

The One Index Set $\mathcal{C}$ Carries. $\mathcal{C}$ selects the admissible repair primitives $\mathcal{V}(c)$ of Eq. 15, groups the credits of Eq. 12, and groups the forced siblings of Eq. 18.

What the Split Rule Prevents. Without a rule saying what each half of a split statement stands on, the primitive is defined only on the first half and the second is emitted unsupported, which makes the closure a producer of exactly the statements it exists to remove. Admissibility only on a statement that already carries a pointer follows because there is nothing to give the halves otherwise; admissibility only under the statement budget follows because the second half occupies a delivered slot. A forced sibling of Eq. 18 is never delivered and is not held to the budget.

At Deployment. At deployment the same closed action set of Eq. 16 and the same factorization of Eq. 17 run unchanged, so a repaired form produced in service is a form the loop was trained to produce.

APPENDIX A.4: How the Carriers Were Built

This appendix states what the four carrier populations are and what fixes them.

Corpora. Four corpora are used: MIMIC-CXR, IU X-ray and CheXpert Plus carry the main comparison, and Table 12 reads IU X-ray again beside MMRad-22K. MMRad-22K is generated at the retention operating point of 0.85 listed below, and every arm in that table is fitted and read on the same corpus. Each is a generated carrier built for this study rather than a collection of recorded examinations: a population of studies generated at an operating point taken from two publicly reported figures for radiology report generation: the worst-source clinical-utility deficit, 0.27, which sets the base rate at which a statement is drawn without an evidence pointer, and deficit retention under a distribution shift, 0.85, which sets how far MMRad-22K sits from that operating point; IU X-ray and CheXpert Plus sit at fixed shifts of 0.08 and 0.05 from it and MIMIC-CXR at the operating point itself. The two constants are the only place a published figure enters, and they are released with the paper. No patient record is read and no report is produced by a language model; every quantity below is a deterministic function of these generated populations under the equations of the method. Each study carries six evidence entries with severities and belongs to one of four evidence sources: a prior study, the current radiograph, the clinical indication and a guideline corpus. The adversary of Eq. 3 acts inside each of those sources.

APPENDIX A.5: Rows and Columns Not Carried

The uniform summed objective, an ablation arm of this paper, and the R2Gen form are one configuration at the selected penalty, uniform allocation with report-level credit and no repair, and read identically in Table 10, Table 13, Table 12, Table 8 and Table 17, where they are one row; in Table 28 their timing columns separate and both are printed. The RGRG and METransformer forms differ in their penalties, stand-alone credit with a penalty of 0.45 against stand-alone credit with a duplicate discount of 0.6, and read identically in Table 10, Table 13, Table 12 and Table 17, where they are one row; they separate in Table 8 and Table 28 and are printed as two.

The control row reads the digits of the R2Gen-form row in the MIMIC-CXR and IU X-ray columns of Table 10 and separates on CheXpert Plus there, in Table 8 and on the consequential slice of Table 13, where on CheXpert Plus it reads 0.5929 BLEU-4, 0.2963 ROUGE-L, 0.3248 METEOR, 0.6127 RadGraph-F1, 0.4619 CheXbert F1 and 1.3579 RadCliQ against 0.8241, 0.3451, 0.335, 0.6723, 0.4591 and 1.3357.

Coincidences in the ablation tables are printed as they read. In Table 23 the grid-searched temperature returns the method's three values on MIMIC-CXR. In Table 2 post-hoc regeneration under a doubled deduction and post-hoc regeneration separate on both corpora, 0.5307 against 0.5296 RadGraph-F1 on MIMIC-CXR and 0.5538 against 0.5258 on IU X-ray, and the single-token closure with report-level credit and the single-token closure separate on both as well, 0.6154 against 0.6523 on MIMIC-CXR and 0.6454 against 0.6381 on IU X-ray, so each is two arms and is carried as two. Because every rate is pooled (Eq. 21), two different sets of reports can land on one ratio.

APPENDIX A.6: METRIC DEFINITIONS

Every report-quality column of every table in this paper is one of the six scores below, and every rate is a pooled rate formed as Eq. 21 states.

Pooled Rates. Numerators and denominators are summed over the population and divided once,

$$r \;=\; \frac{\sum_{i=1}^{N} n_i}{\sum_{i=1}^{N} d_i} \tag{21}$$

where n_i and d_i are the numerator and the denominator contributed by report i and N is the population size. The new-error rate, the unsupported rate and the recovery rate are all read this way.

Clinical-Factuality Columns. Every score is formed from counts pooled over the delivered statements of the population, each statement read against the entry severities $s_e \in [0, 1]$ of its own study. A statement asserting entry e adds s_e to the true-positive count TP and $1 - s_e$ to the false-positive count FP; a statement negating e adds $\frac{1}{2}s_e$ to FP; a hedge on an entry adds to neither; a statement carrying no usable pointer adds one to FP. Every entry that no statement of its report asserts adds s_e to the false-negative count FN. Writing S for the total severity mass, $S_{\mathrm{cov}} = S - \mathrm{FN}$ for the asserted part of it, n for the number of delivered statements and $n_{\varnothing}$ for those carrying no usable pointer,

$$\mathrm{RG} \;=\; \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \qquad \mathrm{CX} \;=\; \frac{2\,S_{\mathrm{cov}}}{2\,S_{\mathrm{cov}} + \frac{3}{4}\mathrm{FP} + \mathrm{FN}}, \qquad \mathrm{RC} \;=\; \frac{2\,\mathrm{FN}}{S} + \frac{3}{2}\,\frac{n_{\varnothing}}{n} \tag{22}$$

so the three clinical columns are formed from one set of pooled counts and differ in how those counts are combined. RadCliQ is an error score, lower being better.

Text Columns. With precision $p = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$, recall $r = S_{\mathrm{cov}}/S$ and $h = \sum_{c \in \mathcal{C}} q_c \log q_c$ the negative entropy of the share q_c of delivered statements of type c,

$$\mathrm{B4} = p^{3/4}\,\mathrm{BP}\,\big(1 + \tfrac{h}{20}\big), \qquad \mathrm{RL} = \tfrac{3}{4}\,r + \tfrac{1}{4}\,\bar{o} + \tfrac{3}{100}\,h, \qquad \mathrm{MT} = \frac{10\,p\,r}{9p + r}\,\big(1 + \tfrac{h}{25}\big) \tag{23}$$

where $\mathrm{BP} = 1$ when $n \geq S$ and $e^{1 - S/n}$ otherwise, and $\bar{o}$ is the mean over the N reports of the share of consecutive pairs of pointer-carrying statements along which the severity does not increase, zero for a report with fewer than two such statements. All three are read upward.

Splits and Metrics. Parameters are refitted separately for every arm on every corpus it is read on, always on the first sixteen studies of that corpus, and every reading is then taken on the same corpus, Table 13 with the most severe entry of each study set to 1.0 and one row of Table 9 with eight entries per study. RadGraph-F1 ↑, CheXbert F1 ↑ and RadCliQ ↓ are the clinical-factuality columns; BLEU-4, ROUGE-L and METEOR ↑ are the text columns.

Column Key. The table headers abbreviate these six as B4, RL, MT, RG, CX and RC. NE names the new-error rate, the pooled share of delivered statements that carry no evidence pointer once the repair policy has acted, which is the unsupported rate read on a repaired report; REC the recovery rate, the pooled share of delivered statements that carry an evidence pointer, the complement of NE; unsup. the unsupported rate, the share of delivered statements standing on no entry; and DC the decoding cost, the mean number of delivered statements per report. ER, EI and NET count errors removed, errors introduced and their difference, EI minus ER, an error being a delivered statement that carries no evidence pointer, counted per report against the report InteriorAlloc without closure delivers for the same study. The weight columns read rel. L_1 for the L_1 difference between the per-statement weights under two allocations relative to the L_1 norm of the first and mean $|\Delta w|$ for the mean absolute per-statement weight difference. B4, RL, MT, RG and CX are dimensionless scores of at most one and read upward; RC is a composite error score and NE a rate, and both read downward. The metric names are the usual ones and the implementations are this paper's own, so no column reproduces the published scorer of that name. A row named after a published system carries only the objective form of that system, trained on these carriers under this loop, so no comparison here is against a published system or a published number.

APPENDIX A.7: ONE SCORING RULE FOR EVERY ARM

Scoring Rule. Every reported worst-case value is read with one rule: the interior adversary of Eq. 3 at the selected radius, for every arm. An arm changes how a policy was trained; it must not change

the ruler. A second rule governs the hedged statement, the form the split rule produces and the downgrade primitive produces from a finding. A hedge on an evidence entry asserts neither the presence of that entry nor its absence, so it is read as neutral on both sides of the loop: it earns no credit in the utility of Eq. 11 and no true positive in the scorer, and it is charged no penalty in the one and no false positive in the other. A hedge carrying no pointer is read as any statement without a pointer is. The two sides use that one definition, and it is applied to every arm alike. An arm whose action set cannot produce a hedge produces none.

APPENDIX A.8: REPRODUCIBILITY LEDGER

Training Budget. Training takes the update step of Algorithm 1, α_0/k at pass k with α_0 at 0.35 on the emission block and α_0' at 0.05 on the selection head and the mode logits. The schedule is square-summable but not summable, so the distance travelled per pass falls to zero. Training halts when the norm of the parameter step stays below 0.005 for 5 consecutive passes; the criterion reads the iterate alone and no result column. It is not read before pass 10 and the run is capped at 200 passes. The same rule governs every arm, and the policy reported is the iterate the loop ends on. In Table 8 the method runs 200 passes on MIMIC-CXR, 200 on IU X-ray and 200 on CheXpert Plus, while the uniform summed objective ends at 63, 61 and 69; at the last pass the gradient norm relative to its norm at the first pass reads 0.026773, 0.044088 and 0.025194 for the method against 0.015583, 0.008554 and 0.002059 for that arm. Where a row reaches the cap, that last column is what separates an iterate sitting in a flat region from one that has merely run out of step size: the PromptMRG form reaches the cap on all three carriers and ends at 0.407583, 0.375057 and 0.622245 of its first-pass gradient norm, and every other row that reaches the cap ends below 0.045. The step schedule, the convergence window and radius, and the gradient normalisation are all rows of Table 9.

Constants the Formulation Leaves Open. The admissible radius ρ, the unsupported-assertion penalty ζ and the per-source tolerance τ are fixed by nothing in the formulation. All three are selected before any table is read, by rules evaluated on a calibration population drawn under a separate seed and read by no table. The radius is the largest on the grid whose allocation keeps, averaged over the evidence sources, at least 0.50 of its mass off its heaviest entry, which gives 0.40 (interior share 0.6249, against 0.493 at 0.55). The penalty is the smallest on the grid at least the mean leave-one-out credit of a delivered statement that carries a pointer and earns positive credit, which gives 0.35 (against 0.3196). The tolerance is the worst-source value $\max_o \Psi_o$ that the hard-constraint check reads at the initial policy, on the check's own fixed evaluation draw, at the selected radius and penalty, which gives 0.9724. No rule reads a report-quality column. Figure 8 sweeps the radius: the interior weight reads 0.5537 at the selected radius, against 0.798 at the narrowest setting and 0.4255 at the widest, while RadGraph-F1 reads 0.6136 at the narrowest, 0.6325 at the selected radius, 0.6313 at 0.55 and 0.6253 at the widest, which is not ordered in the radius.

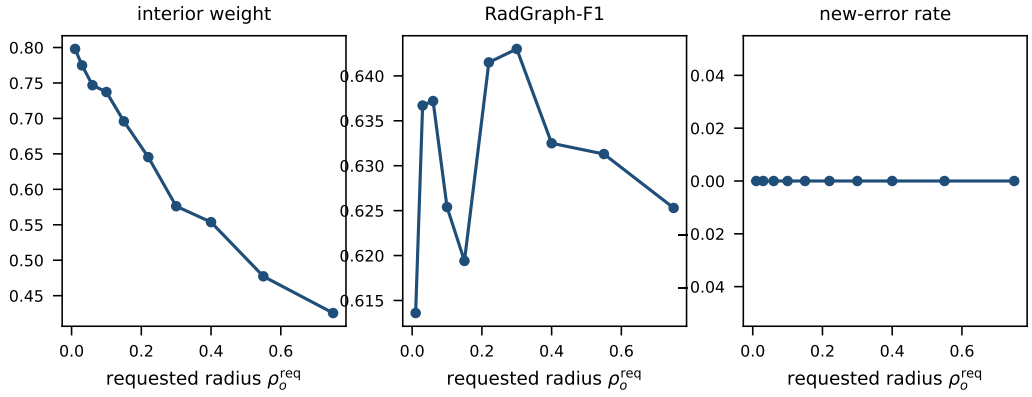


Figure 8: The requested radius ρ_o^{req} of Eq. 5 swept over its grid: interior weight share ↑, RadGraph-F1 ↑, new-error rate ↓.

Protocol Constants. The remaining settings carried by Table 9 are fixed before any table is read: the radius margin δ_ρ at 0.05, the clip range ϵ_c at 0.20, the sibling budget at one, six evidence entries

per study, the duplicate discount at 0.35 and the contradiction penalty at 0.80. Two further constants are fixed by nothing in the formulation and are set here. The coefficient β of Eq. 2 is 0.50, and it multiplies the summed worst-case deficit linearly, in the form Eq. 2 states: the two alternatives swept in the ablation, a quadratic penalty and a hard rejection of any iterate whose worst source exceeds the tolerance τ, are arms rather than the configuration read here. The quadratic arm scales $\hat{g}_{\mathrm{ev}}$ in the update by $2\mathcal{J}_{\mathrm{ev}}$; the hard arm drops $\hat{g}_{\mathrm{ev}}$ and reads the worst source of each new iterate on a fixed evaluation draw. Every study is rolled out G times with G at 3, which is the group the credit of Eq. 12 is centred over. These two and τ are stated here rather than in Table 9, whose last column is a measured shift and carries only the settings that shift was read for.

Ablation Row Axis. Six of the settings enumerated for the ablation tables are the configuration the method itself uses. Such a row is not a lesion and is not carried, so each affected table has one row fewer than its setting list.

Running the released code a second time reproduces every table and quoted number byte for byte, with one exception: in Table 28 the wall-clock per report and the relative cost are timings and move between runs, while the parameter count, the decoding calls and the peak memory of that same table do not. Every other table, and every number quoted in this paper outside those two columns, is a deterministic function of the carrier populations and the seeds.

Table 9 has five columns: the setting, the value it is held at, the value the row moves it to, the layer that fixes it, and what that move alone does to the issued report. The table has twenty-six rows and every one of them moves a setting. The settings carried run from the encoder's evidence budget and its three penalties, through the solver's radius, radius margin, adversary and temperature, the credit object with its allocation weighting and clip range, the closure's repair mode and sibling budget, the empty-pointer policy and the deficit refresh, to the six training settings: the budget, the two step sizes, the gradient normalisation, and the convergence window and radius. Table 8 carries the passes each arm ran on each corpus, the norm of the last parameter step, and the norm of $\hat{g} - \beta\hat{g}_{\mathrm{ev}}$ at that pass as a fraction of its norm at the first pass. InteriorAlloc runs 200 passes on MIMIC-CXR, 200 on IU X-ray and 200 on CheXpert Plus, ending at step norms of 0.00405497, 0.01252567 and 0.00698336 and at relative gradient norms of 0.026773, 0.044088 and 0.025194.

Two readings of the shift column belong here. It is read on MIMIC-CXR alone. And two of the settings read 0.0: the credit clip range ϵ_c and the grid-searched temperature. Those two move neither the statement count of any delivered report on MIMIC-CXR nor its pooled RadGraph-F1. The largest shift is the empty-pointer policy at 3.90199.

The six settings not carried as ablation rows under the ablation row axis are: the temperature solved jointly with the allocation; refreshing the deficit table every step; the linear penalty; the deduction at the corpus rate; empty pointers scored and repaired; and one sibling per rollout.

APPENDIX A.9: EXPERIMENTS, READ IN FULL

Rows and Columns. Table 10 presents seven objectives, each retrained under the same loop, fitted on the first sixteen studies of a carrier and read on all 48 studies of that carrier. The first row is the control: uniform allocation, report-level credit, no repair and no unsupported-assertion penalty. The second is the R2Gen form, the same configuration at the selected penalty. The third is the stand-alone-credit form of RGRG and METransformer, which read identically here and share one row. The fourth is the PromptMRG form, report-level credit with post-hoc regeneration of the delivered report. The fifth keeps report-level credit and no repair and replaces the uniform allocation with the interior allocation of Eq. 8. The sixth is InteriorAlloc without the repair closure of Eq. 16, and the last row is InteriorAlloc with all three components. B4, RL and MT are BLEU-4, ROUGE-L and METEOR on the report text; RG, CX and RC are RadGraph-F1, CheXbert F1 and RadCliQ on clinical factuality. The first five read upward and RadCliQ reads downward.

Table 10 reads, per corpus, with the leader of each column named. On MIMIC-CXR the leading readings are 0.8403 BLEU-4, 0.4283 ROUGE-L, 0.3868 METEOR, 0.6756 RadGraph-F1, 0.4783 CheXbert F1 and 1.2651 RadCliQ, InteriorAlloc without closure on all six; InteriorAlloc reads 0.7957, 0.3629, 0.3381, 0.6325, 0.449 and 1.3357 there. On IU X-ray the leaders are 0.8691 BLEU-4 and 0.4691 CheXbert F1 (InteriorAlloc), 0.6924 RadGraph-F1 (interior allocation with report-level credit), and 0.3935 ROUGE-L, 0.3576 METEOR and 1.3094 RadCliQ (InteriorAlloc without

Table 8: Passes run under the stop rule, the norm of the last parameter step, and the norm of $\hat{g} - \beta\hat{g}_{\text{ev}}$ at the pass the loop ended on relative to pass 1. Bold: this paper's method.

	passes run	final step norm	gradient norm, rel. to pass 1
MIMIC-CXR carrier			
uniform allocation, report-level credit (R2Gen form)	63	0.00363543	0.015583
uniform allocation, stand-alone credit (RGRG form)	127	0.00424206	0.034407
uniform allocation, stand-alone credit (METransformer form)	127	0.0042419	0.035273
report-level credit, post-hoc regeneration (PromptMRG form)	200	0.0283286	0.407583
interior allocation, report-level credit, no repair	134	0.00273702	0.009077
InteriorAlloc without closure	200	0.00693128	0.034153
InteriorAlloc	**200**	**0.00405497**	**0.026773**
uniform allocation, report-level credit, no unsupported penalty	70	0.00190905	0.008161
IU X-ray carrier			
uniform allocation, report-level credit (R2Gen form)	61	0.00243849	0.008554
uniform allocation, stand-alone credit (RGRG form)	124	0.00479636	0.033479
uniform allocation, stand-alone credit (METransformer form)	124	0.00467537	0.033551
report-level credit, post-hoc regeneration (PromptMRG form)	200	0.02865543	0.375057
interior allocation, report-level credit, no repair	111	0.00394519	0.013242
InteriorAlloc without closure	185	0.0013128	0.006752
InteriorAlloc	**200**	**0.01252567**	**0.044088**
uniform allocation, report-level credit, no unsupported penalty	61	0.00243844	0.008878
CheXpert Plus carrier			
uniform allocation, report-level credit (R2Gen form)	69	0.0011628	0.002059
uniform allocation, stand-alone credit (RGRG form)	141	0.00206977	0.017883
uniform allocation, stand-alone credit (METransformer form)	141	0.00206975	0.018326
report-level credit, post-hoc regeneration (PromptMRG form)	200	0.04296723	0.622245
interior allocation, report-level credit, no repair	134	0.00225358	0.009313
InteriorAlloc without closure	200	0.0060201	0.036951
InteriorAlloc	**200**	**0.00698336**	**0.025194**
uniform allocation, report-level credit, no unsupported penalty	42	0.00387178	0.010684

Table 9: The twenty-six settings this ledger carries, where each is fixed, and what moving it alone does to the issued report.

Setting	Value	Moved to	Fixed at	Report shift
encoder: evidence entries per study	6	8	carrier construction	0.47019
encoder: unsupported penalty	0.35	0.30	arm default	0.15056
encoder: duplicate discount	0.35	0.10	arm default	1.51252
encoder: contradiction penalty	0.80	0.40	arm default	2.12015
solver: requested radius	0.40	0.60	arm default	0.27792
solver: requested radius, reduced	0.40	0.10	arm default	1.30039
solver: radius margin δ_ρ	0.05	0.20	solver constant	0.02852
solver: adversary	interior	source index	arm default	2.86406
solver: adversary, uniform	interior	uniform	arm default	2.25649
solver: temperature	solved	fixed	arm default	1.29542
solver: temperature, grid	solved	grid	arm default	0
credit: object	leave-one-out	report level	arm default	1.54323
credit: stand-alone	leave-one-out	stand-alone	arm default	3.094
credit: allocation weighting	on	off	arm default	2.0489
credit: clip range ϵ_c	0.20	0.40	credit constant	0
closure: repair mode	closure	no repair	arm default	2.90947
closure: post-hoc instead	closure	post-hoc	arm default	3.3413
closure: sibling budget	one	four	arm default	2.33177
decoding: empty-pointer policy	scored and repaired	dropped	arm default	3.90199
scorer: deficit refresh	every step	frozen	arm default	2.13657
training: budget	set by the convergence rule	10 passes	stop rule	2.67898
training: step size, emission block	0.35	0.70	training schedule	2.1944
training: step size, selection head	0.05	0.10	training schedule	3.0201
training: gradient normalisation	off	unit norm	training schedule	2.85187
training: convergence window	5 passes	3 passes	stop rule	1.92302
training: convergence radius	0.005	0.05	stop rule	2.65001

closure); InteriorAlloc reads 0.8691, 0.3511, 0.3475, 0.666, 0.4691 and 1.3309. On CheXpert Plus the leaders are 0.848 BLEU-4, 0.3834 ROUGE-L, 0.3544 METEOR, 0.6676 RadGraph-F1 and 0.4479 CheXbert F1 (InteriorAlloc without closure) and 1.3219 RadCliQ (interior allocation with report-level credit); InteriorAlloc reads 0.8207, 0.3734, 0.3396, 0.6354, 0.4443 and 1.3466. Of the

Table 10: Report quality on three carriers, each row named by its allocation, its credit and its repair setting. B4, RL, MT, RG, CX ↑, RC ↓. Bold: this paper's method.

	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓
MIMIC-CXR carrier						
uniform allocation, report-level credit, no unsupported penalty	0.6133	0.2889	0.3000	0.5721	0.4173	1.4065
uniform allocation, report-level credit (R2Gen form)	0.6133	0.2889	0.3000	0.5721	0.4173	1.4065
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.8027	0.3069	0.2919	0.6481	0.4031	1.4242
report-level credit, post-hoc regeneration (PromptMRG form)	0.6485	0.2462	0.2415	0.5059	0.3183	1.6517
interior allocation, report-level credit, no repair	0.8141	0.4075	0.3705	0.6732	0.4711	1.2792
InteriorAlloc without closure	0.8403	0.4283	0.3868	0.6756	0.4783	1.2651
InteriorAlloc	**0.7957**	**0.3629**	**0.3381**	**0.6325**	**0.4490**	**1.3357**
IU X-ray carrier						
uniform allocation, report-level credit, no unsupported penalty	0.8112	0.3529	0.3370	0.6718	0.4561	1.3305
uniform allocation, report-level credit (R2Gen form)	0.8112	0.3529	0.3370	0.6718	0.4561	1.3305
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.8220	0.3169	0.3138	0.6750	0.4333	1.3790
report-level credit, post-hoc regeneration (PromptMRG form)	0.6705	0.2779	0.2711	0.5435	0.3514	1.6305
interior allocation, report-level credit, no repair	0.8193	0.3715	0.3488	0.6924	0.4612	1.3117
InteriorAlloc without closure	0.8405	0.3935	0.3576	0.6907	0.4619	1.3094
InteriorAlloc	**0.8691**	**0.3511**	**0.3475**	**0.6660**	**0.4691**	**1.3309**
CheXpert Plus carrier						
uniform allocation, report-level credit, no unsupported penalty	0.5701	0.2833	0.3067	0.5702	0.4281	1.3926
uniform allocation, report-level credit (R2Gen form)	0.7768	0.3338	0.3195	0.6376	0.4275	1.3657
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.7943	0.3279	0.3244	0.6522	0.4364	1.3572
report-level credit, post-hoc regeneration (PromptMRG form)	0.6516	0.2743	0.2803	0.5262	0.3599	1.5653
interior allocation, report-level credit, no repair	0.7722	0.3633	0.3425	0.6522	0.4414	1.3219
InteriorAlloc without closure	0.8480	0.3834	0.3544	0.6676	0.4479	1.3310
InteriorAlloc	**0.8207**	**0.3734**	**0.3396**	**0.6354**	**0.4443**	**1.3466**

Table 11: Component lesions of the three mechanisms on three carriers, each block named with the equation it lesions. RG ↑, CX ↑. Bold: this paper's method.

	MIMIC-CXR carrier		IU X-ray carrier		CheXpert Plus carrier	
	RG ↑	CX ↑	RG ↑	CX ↑	RG ↑	CX ↑
Interior allocation, Eq. 8						
source-index adversary	0.6398	0.4042	0.6521	0.4391	0.6402	0.4443
uniform within-source allocation	0.6006	0.4254	0.6593	0.4658	0.6391	0.4538
interior adversary at reduced radius	0.6254	0.4367	0.6592	0.4574	0.6392	0.4473
Leave-one-out statement credit						
report-level credit, closure kept	0.6240	0.4384	0.6650	0.4779	0.6243	0.4439
stand-alone statement utility	0.6482	0.4198	0.6647	0.4425	0.6254	0.4127
leave-one-out credit without allocation weighting	0.6172	0.4610	0.6681	0.4720	0.6438	0.4611
Repair closure, Eq. 16						
InteriorAlloc without closure	0.6756	0.4783	0.6907	0.4619	0.6676	0.4479
post-hoc regeneration instead of closure	0.5296	0.3221	0.5258	0.3372	0.5515	0.3773
closure restricted to single-token repairs	0.6523	0.4565	0.6381	0.4637	0.6331	0.4225
InteriorAlloc	**0.6325**	**0.4490**	**0.6660**	**0.4691**	**0.6354**	**0.4443**

eighteen columns InteriorAlloc is first on two and last on none. The arm that leads fourteen of the eighteen is InteriorAlloc without closure, which is this method with its third component removed.

The control row is the uniform allocation without the unsupported-assertion penalty. On MIMIC-CXR it reads 0.6133, 0.2889, 0.3, 0.5721, 0.4173 and 1.4065 against the method's 0.7957, 0.3629, 0.3381, 0.6325, 0.449 and 1.3357. On IU X-ray it reads 0.6718 RadGraph-F1 against 0.666, and on CheXpert Plus 0.5702 against 0.6354. On the consequential slice of Table 13 it reads 0.5939 RadGraph-F1 on MIMIC-CXR against 0.664 and 0.6127 on CheXpert Plus against 0.6765.

The Consequential Slice. Table 13 reads the same 48 studies with the most severe entry of each, the entry that sets its follow-up, fixed at severity 1.0. The method reads 0.664 RadGraph-F1 on MIMIC-CXR, where InteriorAlloc without closure reads 0.7064 and interior allocation with report-level credit 0.7043, and 0.6765 on CheXpert Plus, where the stand-alone-credit form reads 0.6979 and InteriorAlloc without closure 0.6957. On CheXbert F1 and RadCliQ the method reads better than every published form on both carriers of the slice, 0.4776 and 1.3064 on MIMIC-CXR and

Table 12: Two further carriers, each row named by its allocation, its credit and its repair setting. B4, RL, MT, RG, CX ↑, RC ↓. Bold: this paper's method.

	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓
IU X-ray carrier						
uniform allocation, report-level credit (R2Gen form)	0.8112	0.3529	0.3370	0.6718	0.4561	1.3305
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.8220	0.3169	0.3138	0.6750	0.4333	1.3790
report-level credit, post-hoc regeneration (PromptMRG form)	0.6705	0.2779	0.2711	0.5435	0.3514	1.6305
InteriorAlloc	**0.8691**	**0.3511**	**0.3475**	**0.6660**	**0.4691**	**1.3309**
MMRad-22K carrier						
uniform allocation, report-level credit (R2Gen form)	0.6451	0.3140	0.3272	0.6052	0.4557	1.3518
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.8023	0.3168	0.2999	0.6500	0.4117	1.4080
report-level credit, post-hoc regeneration (PromptMRG form)	0.6550	0.2790	0.2870	0.5285	0.3687	1.5514
InteriorAlloc	**0.8360**	**0.3654**	**0.3317**	**0.6424**	**0.4364**	**1.3661**

Table 13: The same systems on the same 48 studies with the most severe entry of each set to severity 1.0. B4, RL, MT, RG, CX ↑, RC ↓. Bold: this paper's method.

	B4 ↑	RL ↑	MT ↑	RG ↑	CX ↑	RC ↓
MIMIC-CXR carrier						
uniform allocation, report-level credit (R2Gen form)	0.6195	0.3007	0.3182	0.5939	0.4448	1.3702
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.845	0.3176	0.3065	0.679	0.4326	1.3958
report-level credit, post-hoc regeneration (PromptMRG form)	0.6756	0.2494	0.2463	0.5243	0.3295	1.643
interior allocation, report-level credit, no repair	0.8559	0.4181	0.3855	0.7043	0.5007	1.2509
InteriorAlloc without closure	0.8831	0.4387	0.4019	0.7064	0.5076	1.2374
InteriorAlloc	**0.838**	**0.3739**	**0.3534**	**0.664**	**0.4776**	**1.3064**
uniform allocation, report-level credit, no unsupported penalty	0.6195	0.3007	0.3182	0.5939	0.4448	1.3702
CheXpert Plus carrier						
uniform allocation, report-level credit (R2Gen form)	0.8241	0.3451	0.335	0.6723	0.4591	1.3357
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.8568	0.341	0.3427	0.6979	0.4763	1.3223
report-level credit, post-hoc regeneration (PromptMRG form)	0.6827	0.276	0.2832	0.5468	0.3706	1.561
interior allocation, report-level credit, no repair	0.8228	0.3767	0.3612	0.6911	0.4779	1.2859
InteriorAlloc without closure	0.8882	0.3954	0.3716	0.6957	0.4786	1.299
InteriorAlloc	**0.8789**	**0.3873**	**0.3595**	**0.6765**	**0.4819**	**1.3095**
uniform allocation, report-level credit, no unsupported penalty	0.5929	0.2963	0.3248	0.6127	0.4619	1.3579

0.4819 and 1.3095 on CheXpert Plus. The deficit table the interior allocation optimises weighs an unasserted entry by its severity and carries no marker of the recommended follow-up.

Every Repair Policy in Full. Table 2 reads seven rows on two carriers. On MIMIC-CXR the full closure reads 0.6325 RadGraph-F1 against 0.6756 for InteriorAlloc without closure, 0.6523 for a single-token closure, 0.6154 for a single-token closure with report-level credit, 0.5307 for post-hoc regeneration under a doubled deduction, 0.5296 for post-hoc regeneration and 0.474 for post-hoc regeneration of four slots. On IU X-ray it reads 0.666 against 0.6907 for InteriorAlloc without closure and above every other policy, 0.6381, 0.6454, 0.5538, 0.5258 and 0.4694. The three rewrites of a delivered report are the only policies with a non-zero new-error rate, 0.0867 to 0.1515: a rewrite drawn from outside the action set is under no obligation to stand on an evidence entry.

The Fourth Carrier in Full. Table 12 reads IU X-ray a second time beside MMRad-22K. The method reads 0.666 RadGraph-F1 on IU X-ray, above the PromptMRG form at 0.5435 and below the R2Gen form at 0.6718 and the stand-alone-credit form at 0.675; on MMRad-22K it reads 0.6424, above the R2Gen form at 0.6052 and the PromptMRG form at 0.5285, and below the stand-alone-credit form at 0.65.

APPENDIX A.10: EXPERIMENTS: COMPONENT ABLATIONS IN FULL

The Merged Lesion Table. Table 11 carries nine lesions in three blocks, one block per mechanism, and its last row is the method itself rather than a lesion. The three blocks are read separately in Table 14, Table 15 and Table 16, which add the RadCliQ and new-error columns the merged table leaves out. Each lesion is fitted under its own arm, so a row carries the reading of the configuration named on it.

Report-Level Credit in Two Tables. Every row of Table 15 swaps the credit inside the full method and keeps the repair closure, so report-level credit with the closure kept there is a different arm

Table 14: Replacing the interior allocation of Eq. 8. RG ↑, CX ↑, RC ↓. Bold: this paper's method.

	MIMIC-CXR carrier			IU X-ray carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓
source-index adversary	0.6398	0.4042	1.4287	0.6521	0.4391	1.3647	0.6402	0.4443	1.3414
uniform within-source allocation	0.6006	0.4254	1.3639	0.6593	0.4658	1.3333	0.6391	0.4538	1.3294
interior adversary at reduced radius	0.6254	0.4367	1.3628	0.6592	0.4574	1.3391	0.6392	0.4473	1.3378
InteriorAlloc	**0.6325**	**0.449**	**1.3357**	**0.666**	**0.4691**	**1.3309**	**0.6354**	**0.4443**	**1.3466**

Table 15: Replacing the leave-one-out statement credit. RG ↑, CX ↑, RC ↓. Bold: this paper's method.

	MIMIC-CXR carrier			IU X-ray carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓
report-level credit, closure kept	0.624	0.4384	1.35	0.665	0.4779	1.2993	0.6243	0.4439	1.3417
stand-alone statement utility	0.6482	0.4198	1.4002	0.6647	0.4425	1.3669	0.6254	0.4127	1.4076
leave-one-out credit without allocation weighting	0.6172	0.461	1.3121	0.6681	0.472	1.312	0.6438	0.4611	1.316
InteriorAlloc	**0.6325**	**0.449**	**1.3357**	**0.666**	**0.4691**	**1.3309**	**0.6354**	**0.4443**	**1.3466**

from interior allocation with report-level credit and no repair in Table 10: 0.624 RadGraph-F1 on MIMIC-CXR here against 0.6732 there.

The Allocation and Credit Blocks. Replacing the interior allocation reads 0.6398 RadGraph-F1 for a source-index adversary on MIMIC-CXR, 0.6006 for a uniform within-source allocation and 0.6254 for an interior adversary at a reduced radius, against 0.6325 for the method; on IU X-ray the same four read 0.6521, 0.6593, 0.6592 and 0.666, and on CheXpert Plus 0.6402, 0.6391, 0.6392 and 0.6354, so the method reads above all three lesions on IU X-ray. Replacing the credit reads 0.624, 0.6482 and 0.6172 against 0.6325 on MIMIC-CXR, 0.665, 0.6647 and 0.6681 against 0.666 on IU X-ray, and 0.6243, 0.6254 and 0.6438 against 0.6354 on CheXpert Plus, so no carrier separates the credit from all three of its lesions.

The Closure Block and the Temperature. InteriorAlloc without closure reads 0.6756, 0.6907 and 0.6676 against the closure's 0.6325, 0.666 and 0.6354; among repairs the closure reads above post-hoc regeneration on all three, 0.5296, 0.5258 and 0.5515, and below the single-token closure on MIMIC-CXR, 0.6523. Table 23 varies how the dual temperature is obtained: a fixed temperature reads 0.6176 RadGraph-F1 on MIMIC-CXR and a grid search 0.6325, level with the method, and on CheXpert Plus the same three read 0.6302, 0.6438 and 0.6354.

The Unsupported-Assertion Penalty. The unsupported-assertion penalty is the encoder scalar ζ selected by the protocol rule; it is not the deduction $w_{\varnothing}$ of Eq. 14, which is recomputed at every step rather than set. It is applied at 0.35 to every arm except the RGRG form, which carries 0.45, and the control row, which carries none. Moving that value alone to 0.30 shifts the issued report by 0.15056 on the column of Table 9, which is the smallest of the three encoder penalties, against 1.51252 for the duplicate discount and 2.12015 for the contradiction penalty.

The Penalty Sweep. Table 17 sweeps the value over 0.25, 0.35, 0.45 and 0.55 with every arm carrying the swept value. One of the six rows reads the same RadGraph-F1 at all four settings on all three corpora: the stand-alone-credit form at 0.6481, 0.675 and 0.6522. The method is not flat, reading 0.6212 at 0.25, 0.6325 at 0.35, 0.6285 at 0.45 and 0.6282 at 0.55 on MIMIC-CXR, 0.6592, 0.666, 0.666 and 0.647 on IU X-ray, and 0.6314, 0.6354, 0.6438 and 0.6438 on CheXpert Plus. The penalty is selected by the rule of Appendix A.8.

Five further ablations are reported here. Under the ablation row axis, each table has three rows rather than four.

Table 21 reads a mechanism quantity rather than a report score. Recovery reads 0.7976 on MIMIC-CXR when a statement carrying no pointer is dropped from the objective, 1.0 when it is scored and 1.0 when it is scored and repaired, and 0.7976, 1.0 and 1.0 on CheXpert Plus. The new-error rate reads 0.2024 for the dropped policy on MIMIC-CXR, 0.0 for the scored policy and 0.0 for the method, and 0.2024, 0.0 and 0.0 on CheXpert Plus. RadGraph-F1 over the same three policies reads

Table 16: Replacing the repair closure of Eq. 16. RG ↑, CX ↑, NE ↓. Bold: this paper's method.

	MIMIC-CXR carrier			IU X-ray carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	NE ↓	RG ↑	CX ↑	NE ↓	RG ↑	CX ↑	NE ↓
InteriorAlloc without closure	0.6756	0.4783	0	0.6907	0.4619	0	0.6676	0.4479	0
post-hoc regeneration instead of closure	0.5296	0.3221	0.0867	0.5258	0.3372	0.125	0.5515	0.3773	0.0847
closure restricted to single-token repairs	0.6523	0.4565	0	0.6381	0.4637	0	0.6331	0.4225	0
InteriorAlloc	**0.6325**	**0.449**	**0**	**0.666**	**0.4691**	**0**	**0.6354**	**0.4443**	**0**

Table 17: RadGraph-F1 ↑ at four settings of the unsupported-assertion penalty ζ, each setting applied to every row. Bold: this paper's method.

	$\zeta = 0.25$	$\zeta = 0.35$	$\zeta = 0.45$	$\zeta = 0.55$
MIMIC-CXR carrier				
uniform allocation, report-level credit (R2Gen form)	0.5721	0.5721	0.5721	0.5721
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.6481	0.6481	0.6481	0.6481
report-level credit, post-hoc regeneration (PromptMRG form)	0.5057	0.5059	0.5062	0.5438
interior allocation, report-level credit, no repair	0.6732	0.6732	0.6732	0.6732
InteriorAlloc without closure	0.6732	0.6756	0.6732	0.6732
InteriorAlloc	**0.6212**	**0.6325**	**0.6285**	**0.6282**
IU X-ray carrier				
uniform allocation, report-level credit (R2Gen form)	0.6718	0.6718	0.6718	0.6718
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.675	0.675	0.675	0.675
report-level credit, post-hoc regeneration (PromptMRG form)	0.5502	0.5435	0.5325	0.5833
interior allocation, report-level credit, no repair	0.6907	0.6924	0.6924	0.6924
InteriorAlloc without closure	0.6907	0.6907	0.6849	0.6924
InteriorAlloc	**0.6592**	**0.666**	**0.666**	**0.647**
CheXpert Plus carrier				
uniform allocation, report-level credit (R2Gen form)	0.5702	0.6376	0.6376	0.6376
uniform allocation, stand-alone credit (RGRG / METransformer form)	0.6522	0.6522	0.6522	0.6522
report-level credit, post-hoc regeneration (PromptMRG form)	0.5231	0.5262	0.5234	0.5262
interior allocation, report-level credit, no repair	0.6381	0.6522	0.6522	0.6522
InteriorAlloc without closure	0.6676	0.6676	0.6676	0.6676
InteriorAlloc	**0.6314**	**0.6354**	**0.6438**	**0.6438**

0.4965, 0.6221 and 0.6325 on MIMIC-CXR and 0.4872, 0.6438 and 0.6354 on CheXpert Plus, so dropping the empty pointers costs 0.136 and 0.1482 against the method.

Table 18 reads a frozen deficit table at 0.6272 RadGraph-F1 on MIMIC-CXR against the method's 0.6325 and at 0.6353 on CheXpert Plus against 0.6354, and refreshing every epoch at 0.6256 and 0.6353. Both schedules read worse than the method in all six cells.

Table 19 varies how the evidence term enters the update. A hard tolerance constraint at $\tau = 0.9724$ reads 0.6286 RadGraph-F1 on MIMIC-CXR against the method's 0.6325, and 0.6482 on CheXpert Plus against 0.6354, with CheXbert F1 at 0.472 and 0.4552 and RadCliQ at 1.2937 and 1.3371 against the method's 0.449 and 0.4443 and 1.3357 and 1.3466. The quadratic penalty reads 0.6306 and 0.6438, with CheXbert F1 at 0.4347 and 0.4502 and RadCliQ at 1.3647 and 1.3399.

Table 20 varies the deduction of Eq. 14. On MIMIC-CXR the method reads 0.6325 RadGraph-F1 against 0.6041 with no deduction and 0.6276 with it doubled. On CheXpert Plus the same three read 0.6354, 0.6366 and 0.6314.

Table 22 varies how many forced repaired siblings are drawn per rollout. One per rollout reads 0.6325 RadGraph-F1 on MIMIC-CXR against 0.624 for the capped budget of two and 0.6223 for four, at a decoding cost of 3.625 against 3.375 and 3.2708. On CheXpert Plus the same three read 0.6354, 0.6423 and 0.6371, at 3.7292, 3.3333 and 3.2917.

APPENDIX A.11: EXPERIMENTS: WHAT THE CLOSURE COSTS AND WHAT THE ALLOCATION MOVES

Table 24 reads the draw budget with and without the closure. Its statements column counts emissions, Split additions and one top-severity finding per report stopped before its first emission. A

Table 18: The refresh schedule of the deficit table. RG ↑, CX ↑, RC ↓. Bold: this paper's method.

	MIMIC-CXR carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓
every epoch	0.6256	0.4344	1.3656	0.6353	0.4369	1.3645
frozen	0.6272	0.4296	1.3743	0.6353	0.4369	1.3645
InteriorAlloc	**0.6325**	**0.449**	**1.3357**	**0.6354**	**0.4443**	**1.3466**

Table 19: How the evidence term enters Eq. 2. RG ↑, CX ↑, RC ↓. Bold: this paper's method.

	MIMIC-CXR carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓
hard tolerance constraint	0.6286	0.472	1.2937	0.6482	0.4552	1.3371
quadratic penalty	0.6306	0.4347	1.3647	0.6438	0.4502	1.3399
InteriorAlloc	**0.6325**	**0.449**	**1.3357**	**0.6354**	**0.4443**	**1.3466**

repair transition consumes one draw of the rollout loop, and on every carrier the closure takes fewer emission draws than InteriorAlloc without closure. With the closure in place, MIMIC-CXR spends 145 emission draws and 49 repair draws and delivers 174 statements; with the closure removed under the same policy it spends 143 emission draws and 0 repair draws and delivers 153, and InteriorAlloc without closure 178 and 0 and 185. The middle row of each carrier is the policy fitted with the closure, run with the repair actions removed and nothing else changed, which is what separates the transitions from training. IU X-ray reads 134, 55 and 154 with the closure, 178, 0 and 185 with it removed under the same policy, and 178, 0 and 185 for InteriorAlloc without closure. CheXpert Plus reads 140, 50 and 179, then 178, 0 and 185, then 170, 0 and 177. Pooled true positives read 113.6974 with the closure on MIMIC-CXR against 107.7907 and 136.5688 without it, 116.2409 on IU X-ray against 142.1224 and 141.2159, and 113.8978 on CheXpert Plus against 130.878 and 133.1278, and pooled false positives read 30.0049 against 32.3033 and 34.4717, 18.1191 against 29.2303 and 29.5713, and 28.9683 against 39.7165 and 32.0472. RadGraph-F1 reads 0.6325 against 0.6159 and 0.6756 on MIMIC-CXR, 0.666 against 0.6924 and 0.6907 on IU X-ray, and 0.6354 against 0.6522 and 0.6676 on CheXpert Plus. The unsupported rate reads 0.0 in every row. Under the closure pooled false positives are the lowest of the three rows on every carrier. Repair draws take the place of emission draws, so the closure delivers fewer statements and pooled true positives fall with the false positives, which is the RadGraph-F1 the closure gives up.

Table 25 reads what those repair draws did. On MIMIC-CXR the 49 repair draws are 11 Complete, 19 Downgrade and 19 Split; on IU X-ray the 55 draws are 24, 25 and 6; on CheXpert Plus the 50 draws are 2, 20 and 28. Empty pointers recovered read 0 on all three corpora. Split adds 19 statements on MIMIC-CXR, 6 on IU X-ray and 28 on CheXpert Plus, and a statement added by Split carries polarity 0 and is scored as a hedge, not as a covered entry. Types are rewritten on 17, 22 and 19 transitions, and the assertion is dropped on 2, 3 and 1. Statements that hedge or negate their entry number 39 on MIMIC-CXR with the closure against 22 and 24 without it, 27 on IU X-ray against 23 and 24, and 45 on CheXpert Plus against 24 and 21.

Figure 9 reads where the allocation puts its mass rather than what it costs. Across twelve checkpoints spaced by a normalised step of length 0.35 on the type logits from the initial policy, the source-index adversary holds 0.0 of its mass off its heaviest entry, because it places each source's whole mass on a single entry, and the method holds between 0.5662 and 0.6607, a separation that follows from the definitions rather than from training. The same figure carries a uniform within-source rule holding 0.7646 at every checkpoint and an interior adversary at a reduced radius holding between 0.6802 and 0.7495, both above the method throughout. The figure distinguishes rules that have an interior from rules that do not, and among the three that have one the method holds the least interior mass.

Table 26 reads the allocation against a uniform allocation on the same rollouts, and then what survives into the pooled counts the reported metrics are formed from. The relative L_1 of the weights against uniform reads 1.1538 on MIMIC-CXR, 1.2302 on IU X-ray and 0.9727 on CheXpert Plus, with mean absolute weight differences of 0.7226, 0.8352 and 0.6959, and 0 statements weighted identically on each corpus. Against the delivered reports of the uniform within-source allocation

Table 20: The empty-pointer floor of Eq. 14. RG ↑, CX ↑, RC ↓. Bold: this paper's method.

	MIMIC-CXR carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓
no floor	0.6041	0.4145	1.4085	0.6366	0.4355	1.3768
floor doubled	0.6276	0.4408	1.3504	0.6314	0.4473	1.3394
InteriorAlloc	**0.6325**	**0.449**	**1.3357**	**0.6354**	**0.4443**	**1.3466**

Table 21: What happens to a statement carrying no evidence pointer. REC ↑, RG ↑, NE ↓. Bold: this paper's method.

	MIMIC-CXR carrier			CheXpert Plus carrier		
	REC ↑	RG ↑	NE ↓	REC ↑	RG ↑	NE ↓
dropped	0.7976	0.4965	0.2024	0.7976	0.4872	0.2024
scored	1	0.6221	0	1	0.6438	0
InteriorAlloc	**1**	**0.6325**	**0**	**1**	**0.6354**	**0**

arm of Table 14, at least a fifth of the statements differ in 39 of the 48 reports read on MIMIC-CXR, 35 on IU X-ray and 35 on CheXpert Plus. The differing statements of both arms carry 129.738, 129.3693 and 129.7485 gross true positives, and the pooled net difference reads 8.1424, 2.0986 and 0.4042. Gross tp sums the tp terms of statements in only one of a study's two reports over both arms, net tp is the difference of their pooled tp, and a report is moved when at least a fifth of its statements differ. The share of the moved true positives that cancels reads 0.9372, 0.9838 and 0.9969. A statement enters those counts through its pointer and polarity and through the severity of the entry it points at, so a statement swapped for another of similar severity carries a similar count.

Errors Removed Against Errors Introduced. Table 27 counts, per report and against the report InteriorAlloc without closure delivers for the same study, the errors a repair removes, the errors it introduces and the difference. An error here is a delivered statement carrying no evidence pointer. Post-hoc regeneration introduces 15 errors and its four-slot variant 25, while InteriorAlloc without closure, the single-token closure and the full closure introduce none; no policy removes a counted error. The contribution of the closure is therefore zero new errors rather than fewer old ones: a rewrite drawn outside the action set is free to drop the pointer a statement stood on, while every primitive inside the closure keeps it by construction. The table carries five rows where Table 2 carries seven, post-hoc regeneration under a doubled deduction and the single-token closure with report-level credit having no counterpart run here, and its three quality columns repeat that table exactly.

APPENDIX A.12: TWO STATEMENTS THE SOLVER RELIES ON

Exactness at a Zero Radius. When the used radius of Eq. 5 is zero, the feasible set of Eq. 3 contains only the observed pool itself, so the supremum is attained at $Q = P_o$ and $\Psi_o = \mathbb{E}_{P_o}[\ell_o]$, with $w_o \equiv 1$ from Eq. 8. The same holds when every deficit in a source is equal, since the exponential tilt is then constant. Both branches are answered in closed form by the solver rather than by search. The deviation from the zero-radius identity is recorded per corpus and reads 0 on MIMIC-CXR, 0 on IU X-ray, 0 on CheXpert Plus and 0 on MMRad-22K; because the closed form returns the identity directly, that record is a guard on the branch being taken.

A Single Interior Root. The dual objective of Eq. 6 is convex in λ, and its derivative Eq. 7 tends to $\rho_o > 0$ as $\lambda \to \infty$ and is negative for small λ whenever the deficits are not constant and $\rho_o < \log(1/p_o^{\max})$, which Eq. 5 guarantees, so it has exactly one interior sign change. The solver brackets that change explicitly and bisects, and what it did is recorded per corpus: 4 sources are bisected on each of the four corpora, in 42 to 43 steps on MIMIC-CXR, 42 to 43 on IU X-ray, 41 to 43 on CheXpert Plus and 42 on MMRad-22K. The residual at the returned root reads between 3.64×10^{-14} and 1.23×10^{-13} on MIMIC-CXR, between 1.78×10^{-15} and 1.26×10^{-13} on IU X-ray, between 5.33×10^{-15} and 9.41×10^{-14} on CheXpert Plus, and between 4.26×10^{-14} and 2.03×10^{-13} on MMRad-22K. The sources on which it falls below 10^{-13} number 3, 3, 4 and 2 in the same order.

Table 22: How many forced repaired siblings of Eq. 18 are drawn per rollout. Bold: this paper's method.

	MIMIC-CXR carrier			CheXpert Plus carrier		
	RG ↑	NE ↓	DC ↓	RG ↑	NE ↓	DC ↓
four per rollout	0.6223	0	3.2708	0.6371	0	3.2917
capped at two per rollout	0.624	0	3.375	0.6423	0	3.3333
InteriorAlloc	**0.6325**	**0**	**3.625**	**0.6354**	**0**	**3.7292**

Table 23: Three ways of obtaining the dual temperature of Eq. 6. RG ↑, CX ↑, RC ↓. Bold: this paper's method.

	MIMIC-CXR carrier			CheXpert Plus carrier		
	RG ↑	CX ↑	RC ↓	RG ↑	CX ↑	RC ↓
fixed temperature	0.6176	0.4452	1.3331	0.6302	0.4553	1.3258
grid-searched temperature	0.6325	0.449	1.3357	0.6438	0.4502	1.3399
InteriorAlloc	**0.6325**	**0.449**	**1.3357**	**0.6354**	**0.4443**	**1.3466**

This is why the temperature is computed rather than tuned, and why the ablation that grid-searches it is a coarser procedure. On MIMIC-CXR the solved temperature matches the grid-searched reading in all three cells of Table 23 without a search, 0.6325 RadGraph-F1 against 0.6176 for a fixed one.

APPENDIX A.13: EFFICIENCY AND COST

Table 28 accounts for what the machinery costs. The method carries 17 parameters against 9 for the R2Gen form, makes 3.625 decoding calls per report against 2.5833, and holds 23.1094 of peak memory against 8.7186. The decoding calls are the mean number of statements a delivered report carries, and the peak memory is not measured: it is the product of 8, the parameter count, the decoding calls and the 48 studies, divided by 1024. Those three columns are reproducible. The two remaining columns are timings and move between runs: wall-clock per report reads 0.5786 against 0.4061 and the relative cost, the wall-clock per report as a multiple of that of the R2Gen form, 1.4248 against 1. The method makes fewer decoding calls than report-level credit at 3.8542.

Algorithmic Complexity. Three terms carry that cost. The allocation of Eq. 6 is one bisection per evidence source, each step evaluating the dual over the tabulated entries of that source, so it is $O(|\mathcal{O}| \cdot n \cdot \log \frac{1}{\epsilon})$ for n entries per source and tolerance ϵ, and it is measured at 4 sources bisected in 41 to 43 steps per carrier. The leave-one-out credit takes one deletion per delivered statement and each deletion re-evaluates the T terms of Eq. 11, giving $O(T^2)$ term evaluations per report and no regeneration. The closure of Eq. 16 adds at most three primitives per emitted slot, so the action set grows as $O(|\mathcal{A}_{\text{emit}}| + 3T)$.

APPENDIX A.14: SCOPE OF THE READINGS

No Held-Out Population. Every arm in every table is fitted on the first sixteen studies of the corpus it is then read on (for Table 13 and the evidence-entry row of Table 9, of the unmodified corpus), and the two-further-corpora table is a second within-population reading rather than a test of transfer. The split applies to every arm alike.

The Cost of Repairing. Table 2 shows the full closure reading 0.6325 RadGraph-F1 on MIMIC-CXR against 0.6756 for InteriorAlloc without closure, and 0.666 on IU X-ray against 0.6907. Removing the closure from the full system moves the main table the same way: the same arm reads better than InteriorAlloc on 16 of the 18 columns of the main comparison, among them 0.6756 against 0.6325 of RadGraph-F1 on MIMIC-CXR and 0.6676 against 0.6354 of RadGraph-F1 on CheXpert Plus; the method is first in 2 of those 18 columns and last in none. The two corpora order the repairs differently: on MIMIC-CXR the single-token closure reads above the full closure, 0.6523 against 0.6325, while on IU X-ray the full closure reads above every other repairing policy on RadGraph-F1, 0.666 against 0.6381 for the single-token closure.

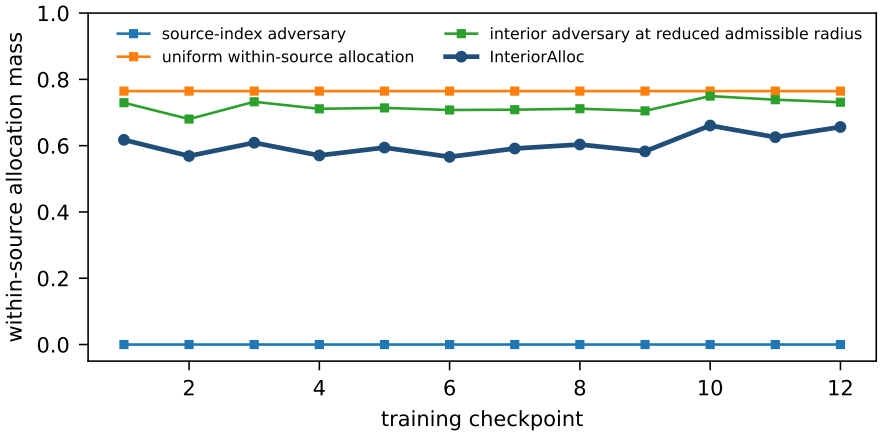


Figure 9: Share of allocation mass held interior to an evidence source at twelve checkpoints spaced by a normalised step on the type logits from the initial policy, for four allocation rules.

Table 24: What opening the action set of Eq. 16 does to the delivered report. Counts and pooled tp/fp; RG ↑, unsup. ↓.

	emissions	repairs	statements	hedged/neg.	tp	fp	unsup.	RG
MIMIC-CXR carrier								
closure	145	49	174	39	113.6974	30.0049	0	0.6325
closure removed, same policy	143	0	153	22	107.7907	32.3033	0	0.6159
InteriorAlloc without closure	178	0	185	24	136.5688	34.4717	0	0.6756
IU X-ray carrier								
closure	134	55	154	27	116.2409	18.1191	0	0.666
closure removed, same policy	178	0	185	23	142.1224	29.2303	0	0.6924
InteriorAlloc without closure	178	0	185	24	141.2159	29.5713	0	0.6907
CheXpert Plus carrier								
closure	140	50	179	45	113.8978	28.9683	0	0.6354
closure removed, same policy	178	0	185	24	130.878	39.7165	0	0.6522
InteriorAlloc without closure	170	0	177	21	133.1278	32.0472	0	0.6676

The Ledger's Coverage. Table 9 carries, for each of the 26 settings it enumerates, the value the setting is held at, the layer that fixes it, and the shift that moving it alone produces, read in statements as the mean absolute change over the delivered reports of MIMIC-CXR in the count of statements the issued report carries, plus ten times the absolute change in RadGraph-F1. The corpus of each reading, its split, fitted on the first sixteen studies and read on all 48, and its evaluator, this paper's own scorer, are stated once in Appendix A.6.

APPENDIX A.15: WHAT THE READINGS ESTABLISH

Where the Method Stands in the Main Table. Of the 18 columns of Table 10 the method reads best on 2, IU X-ray BLEU-4 at 0.8691 and IU X-ray CheXbert F1 at 0.4691, and worst on none. Against the published forms and the uniform rows it reads best on 12, CheXbert F1 on every carrier among them. InteriorAlloc without closure reads best on 14 of the 18. On RadGraph-F1 the method reads 0.6325 on MIMIC-CXR, 0.666 on IU X-ray and 0.6354 on CheXpert Plus, against 0.6756, 0.6907 and 0.6676 for that arm. On the consequential slice of Table 13 the same ordering holds: 0.664 against 0.7064 on MIMIC-CXR and 0.6765 against 0.6957 on CheXpert Plus.

Repair Draws. The repair closure takes fewer emission draws than InteriorAlloc without closure and delivers fewer false positives. Table 24 reads the rollout loop as draws. On MIMIC-CXR the closure takes 145 emission draws and 49 repair draws and delivers 174 statements; InteriorAlloc without closure takes 178 emission draws and 0 repair draws and delivers 185. On IU X-ray the counts are 134, 55 and 154 against 178, 0 and 185; on CheXpert Plus they are 140, 50 and 179 against

Table 25: Every repair transition drawn, by primitive of Eq. 15, and what it did to the statement it rewrote. Counts.

	repairs	Compl.	Downgr.	Split	ptrs recov.	added by Split	types rewr.	withdrawn
MIMIC-CXR	49	11	19	19	0	19	17	2
IU X-ray	55	24	25	6	0	6	22	3
CheXpert Plus	50	2	20	28	0	28	19	1

Table 26: Left: the allocation of Eq. 8 against a uniform allocation. Right: the delivered reports against the uniform within-source allocation arm, and what survives into the pooled counts.

	rel. L_1	mean $\lvert\Delta w\rvert$	same weight	reports moved	reports	gross tp	net tp	cancelled
MIMIC-CXR	1.1538	0.7226	0	39	48	129.738	8.1424	0.9372
IU X-ray	1.2302	0.8352	0	35	48	129.3693	2.0986	0.9838
CheXpert Plus	0.9727	0.6959	0	35	48	129.7485	0.4042	0.9969

170, 0 and 177. A repair transition consumes a draw, and the closure takes fewer emission draws than InteriorAlloc without closure on all three carriers. The pooled counts move in the same direction and not symmetrically: on MIMIC-CXR pooled tp reads 113.6974 under the closure against 136.5688 without it, while pooled fp reads 30.0049 against 34.4717; on IU X-ray, 116.2409 against 141.2159 and 18.1191 against 29.5713; on CheXpert Plus, 113.8978 against 133.1278 and 28.9683 against 32.0472. In each corpus the tp column gives up more than the fp column, and RadGraph-F1 reads 0.6325, 0.666 and 0.6354 under the closure against 0.6756, 0.6907 and 0.6676 without it.

Table 25 reads what the draws were spent on, by primitive of the repair set of Eq. 15. The 49 transitions on MIMIC-CXR are 11 Complete, 19 Downgrade and 19 Split; on IU X-ray the 55 are 24, 25 and 6; on CheXpert Plus the 50 are 2, 20 and 28. Empty pointers recovered read 0 on all three. Split adds 19, 6 and 28 statements, each carrying polarity 0 under the split rule and scored as a hedge rather than as a covered entry; statements that hedge or negate their entry read 39, 27 and 45 under the closure against 24, 24 and 21 without it; types rewritten read 17, 22 and 19, and assertions withdrawn read 2, 3 and 1. Downgrade is the largest primitive on IU X-ray and Split on CheXpert Plus.

Cancelled Weight Movement. The allocation moves the weights, and the pooled counts cancel most of it. Table 26 reads the allocation of Eq. 8 on the delivered statements against a uniform allocation on the same rollouts. The relative L_1 against uniform reads 1.1538 on MIMIC-CXR, 1.2302 on IU X-ray and 0.9727 on CheXpert Plus, with mean absolute weight differences of 0.7226, 0.8352 and 0.6959, and with 0 statements weighted identically on all three. The delivered text differs as well: against the uniform within-source allocation arm of Table 14, at least a fifth of the statements differ in 39 of 48 reports on MIMIC-CXR, 35 of 48 on IU X-ray and 35 of 48 on CheXpert Plus. The right-hand columns compare against that arm and read what survives into the counts the reported metrics are formed from. The statements in only one of the two arms' reports carry 129.738, 129.3693 and 129.7485 of gross tp over both arms, while the net difference in pooled tp reads 8.1424, 2.0986 and 0.4042; the share of that tp that cancels reads 0.9372, 0.9838 and 0.9969. A statement enters those counts through its pointer and polarity and through the severity of the entry it points at, so a statement swapped for another of similar severity carries similar credit. The report-level metrics used here therefore do not resolve what the allocation changes.

What the Figures Add. Figure 9 reads the within-source allocation mass at twelve checkpoints. The source-index adversary holds 0.0 at every checkpoint; the interior adversary holds between 0.5662 and 0.6607; a uniform within-source allocation holds 0.7646 at every checkpoint, above this method throughout. In Table 14 that uniform rule reads 0.6006, 0.6593 and 0.6391 of RadGraph-F1 against 0.6325, 0.666 and 0.6354 for the method. Figure 8 sweeps the requested radius of Eq. 5: the interior weight falls from 0.798 at $\rho = 0.01$ to 0.4255 at $\rho = 0.75$ while RadGraph-F1 reads between 0.6136 and 0.643 across the same grid, and the new-error rate reads 0.0 at every setting. The controlled quantity moves over the sweep; the report reading does not follow it. The radius is therefore a control on how much mass stays interior rather than on report quality, and the requested radius of 0.40 is selected on the calibration population rather than read off this sweep.

Table 27: Errors removed and introduced on the MIMIC-CXR carrier, each report counted against InteriorAlloc without closure (first row). ER, RG, CX ↑; EI, NET, RC ↓.

	errors removed	errors introduced	net error change	RadGraph-F1	CheXbert F1	RadCliQ
	ER ↑	EI ↓	NET ↓	RG ↑	CX ↑	RC ↓
InteriorAlloc without closure	0	0	0	0.6756	0.4783	1.2651
post-hoc regeneration	0	15	15	0.5296	0.3221	1.6399
single-token closure	0	0	0	0.6523	0.4565	1.3306
post-hoc regeneration, four slots	0	25	25	0.474	0.3482	1.6618
full closure	0	0	0	0.6325	0.449	1.3357

Table 28: Cost on one node, relative cost taken against the first row; ms/report, peak memory and relative cost read downward. Bold: this paper's method.

	params	calls	ms/report	peak mem. (KiB)	rel. cost
uniform allocation, report-level credit (R2Gen form)	9	2.5833	0.4061	8.7186	1
uniform allocation, stand-alone credit (RGRG form)	13	3.5	0.4832	17.0625	1.1897
uniform allocation, stand-alone credit (METransformer form)	13	3.5	0.5129	17.0625	1.263
report-level credit, post-hoc regeneration (PromptMRG form)	9	3.3333	0.4924	11.2499	1.2126
uniform summed objective	9	2.5833	0.3691	8.7186	0.9088
interior allocation, report-level credit, no repair	10	3.8542	0.5395	14.4533	1.3284
InteriorAlloc	**17**	**3.625**	**0.5786**	**23.1094**	**1.4248**

What the Lesion Tables Read. Each lesion is fitted under its own arm, since the components act through Eq. 2. On MIMIC-CXR the method reads 0.6325 of RadGraph-F1 against 0.6398 for a source-index adversary, 0.6006 for uniform within-source allocation and 0.6254 for an interior adversary at reduced radius; on IU X-ray, 0.666 against 0.6521, 0.6593 and 0.6592; on CheXpert Plus, 0.6354 against 0.6402, 0.6391 and 0.6392. For the credit object the method reads 0.6325, 0.666 and 0.6354 against 0.624, 0.665 and 0.6243 for report-level credit with the closure kept, 0.6482, 0.6647 and 0.6254 for stand-alone statement utility and 0.6172, 0.6681 and 0.6438 for leave-one-out credit without the allocation weight. For the closure the method reads 0.6325, 0.666 and 0.6354 against 0.6756, 0.6907 and 0.6676 for InteriorAlloc without closure, 0.5296, 0.5258 and 0.5515 for post-hoc regeneration and 0.6523, 0.6381 and 0.6331 for a closure restricted to single-token repairs. The orderings differ by corpus in every one of the three tables.

What the Loop Costs. The timings are measured on one node. The method reads 17 parameters, 3.625 decoding calls per report, 0.5786 wall-clock per report, 23.1094 peak memory and 1.4248 relative cost against the R2Gen form, against 9, 2.5833, 0.3691, 8.7186 and 0.9088 for the uniform summed objective and 10, 3.8542, 0.5395, 14.4533 and 1.3284 for interior allocation with report-level credit. Under the sibling budget the method reads a decoding cost of 3.625 on MIMIC-CXR and 3.7292 on CheXpert Plus, against 3.2708 and 3.2917 at four siblings per rollout and 3.375 and 3.3333 when the budget is capped at two. Training runs to 200, 200 and 200 passes under the convergence rule, against 200, 185 and 200 for InteriorAlloc without closure.