This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material for

When to Request and Whether to Trust: A Two-Decision Imaging Agent for Pleural Effusion Assessment

Yehyun Ko, Junah Jung, Chang Min Park, and Dongheon Lee
Authors and affiliations are identical to those in the main paper.

S1. Objective

This supplementary analysis separates the contribution of lateral-view content from the effect of invoking a second re-reading pathway. We did not train a new request policy. Instead, we froze the learned request policy, including its validation-selected threshold, and applied the same request mask to both test conditions.

For a test study i with frozen request indicator R_i, we defined

$$P_{\text{lateral},i} = \begin{cases} M_{3,i}, & R_i = 1, \\ M_{0,i}, & R_i = 0, \end{cases} \qquad P_{\text{duplicate},i} = \begin{cases} M_{2,i}, & R_i = 1, \\ M_{0,i}, & R_i = 0. \end{cases} \tag{S1}$$

Here, M0 denotes frontal-only inference, M2 denotes re-reading with a duplicated frontal image, and M3 denotes re-reading with the paired lateral image.

$$\Delta_{\text{lateral-specific}} = \text{BA}(P_{\text{lateral}}) - \text{BA}(P_{\text{duplicate}}). \tag{S2}$$

S2. Evaluation protocol

The frozen request mask was applied to the same subject-disjoint test cohort for both arms. The request mask was not changed between conditions, and no policy retraining, threshold retuning, or label-based request selection was performed. Studies with $R_i = 0$ retained the original frontal-only prediction in both arms. For studies with $R_i = 1$, the duplicated-frontal condition used the identical frontal image twice, whereas the lateral condition used the paired lateral image from the same study.

The frozen policy was defined on the main paper's 4,956-study explicit-label test cohort and selected 396 studies (7.99%). For the paired M2–M3 endpoint, three studies with failed duplicated-frontal inference were excluded, leaving 4,953 common complete studies and 395 requested studies. Balanced accuracy (BA) is the mean of sensitivity and specificity. Confidence intervals were obtained with 2,000 subject-level cluster bootstrap replicates using seed 42. The paired confidence interval for the lateral-specific contrast was computed from the same bootstrap replicate in both arms; Parser and inference failures were tracked separately and were excluded only from the paired complete-case endpoint.

S3. Primary frozen-policy comparison

Across the common complete test cohort, the frozen-policy paired-lateral arm achieved a BA of 0.875, compared with 0.869 when the requested lateral image was replaced by a duplicated frontal image. The estimated lateral-specific increment was therefore +0.006 (95% CI: [0.002, 0.010]; Table S1). This contrast is the primary endpoint of this supplementary analysis because it preserves the frozen request decision while changing only the content of the second image for requested studies.

Table S1. Primary frozen-policy comparison on the common complete test cohort ($N = 4{,}953$; 395 requested studies). The frozen request mask is identical across rows, and non-requested studies retain M0. ΔBA is relative to the frontal-only reference except in the final paired-contrast row.

Policy condition	Input for requested studies	BA	ΔBA	95% CI
Frontal-only reference	frontal only	0.867	0.000	[0.856, 0.877]
Frozen policy + duplicated frontal	frontal + duplicated frontal	0.869	+0.002	[0.858, 0.879]
Frozen policy + paired lateral	frontal + paired lateral	0.875	+0.008	[0.864, 0.885]
Paired-lateral − duplicated-frontal contrast	paired lateral − duplicated frontal	–	+0.006	[0.002, 0.010]

Thus, the policy-level improvement is not explained entirely by invoking a second re-reading pathway; a small additional contribution is associated with lateral-view content in this frozen-policy comparison. This analysis complements, rather than replaces, the paper's broader counterfactual audit.

S4. Requested-subset descriptive analysis

For completeness, Table S2 reports the 395 requested studies with valid M2 and M3 predictions. This analysis is descriptive rather than the primary endpoint because it excludes non-requested studies, which remain at M0 in the deployed policy, and the three duplicated-frontal inference failures.

Table S2. Descriptive comparison of the 395 studies requested by the frozen request policy. ΔBA is relative to $M0$ within the requested subset, except in the final paired-contrast row.

Requested-study condition	BA	ΔBA vs. M0	95% CI
M0: frontal only	0.500	0.000	[0.500, 0.500]
M2: frontal + duplicated frontal	0.570	+0.070	[0.530, 0.610]
M3: frontal + paired lateral	0.673	+0.173	[0.627, 0.718]
M3 − M2 paired contrast	–	+0.103	[0.054, 0.155]

The requested-subset contrast is larger than the full-cohort policy contrast because it is restricted to studies selected for a second re-reading. It should therefore not be interpreted as the deployed policy effect. The common-cohort contrast in Table S1 remains the primary result.