NOISE-AWARE DEMONSTRATION SELECTION FOR FAIRNESS-SENSITIVE IN-CONTEXT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot prediction with large language models is often treated as a retrieval problem where the goal is to choose relevant demonstrations. This overlooks a key failure mode in fairness-sensitive settings because same-group, different-group, and desensitized demonstrations can all carry noisy labels. We study the relationship between label noise, group identity, and demonstration selection for in-context prediction, with sex and race as sensitive attributes and prompt-only access to black-box large language models. Regression analyses across two prediction tasks quantify how demonstration label noise interacts with group concordance, desensitization, prompt length, and query-group identity. A central finding is that sensitive-attribute concordance does not protect a prompt from label noise (i.e., same-group demonstrations with noisy labels are still associated with reduced target-label correctness). Motivated by these findings, we introduce Fairness- and Noise-Aware Demonstration Selection (FANDS), which combines estimated noise risk with group relation, similarity, and demonstration-order information to improve few-shot prediction. At selection time, a surrogate model estimates the expected performance of candidate demonstration sets without requiring internal large language model scores or repeated exhaustive prompting. Held-out evaluations show that FANDS improves overall predictive performance and fairness metrics relative to baselines. These results show that demonstration label noise is an important factor in fairness-sensitive in-context prediction and that accounting for both noise and group context can improve few-shot predictive performance and fairness.

1 INTRODUCTION

Large language models (LLMs) can adapt to new prediction tasks from a short instruction and a small set of demonstrations, without updating model parameters (Brown et al., 2020). Their predictions, however, depend on which demonstrations are selected, how they are presented, and their order (Lu et al., 2022; Rubin et al., 2022). These choices especially matter in medical in-context prediction, where demonstrations may contain noisy labels, sensitive-attribute cues, and similar cases. In medical prediction, strong average performance can coexist with disparities across population subgroups (Seyyed-Kalantari et al., 2020; Zhang et al., 2022). This makes demonstration selection a fairness-relevant design choice rather than only an accuracy optimization problem.

Prior work has studied these aspects of demonstration selection separately. Retrieval and learned-selection methods select examples to improve average task performance under a fixed evaluation protocol (Liu et al., 2022; Wu et al., 2024). Fairness-aware prompting methods examine how prompt composition, representation, or predictive bias affects group-conditioned outcomes (Hu et al., 2024; Aguirre et al., 2024). Noisy-demonstration studies show that incorrect or irrelevant in-context examples can harm performance, although the magnitude depends on the task and modality (Gao et al., 2024; Rieff et al., 2025). Less understood is how label noise interacts with sensitive-attribute presentation. A demonstration may closely resemble the query and belong to the same demographic group, yet display an incorrect label. Similarly, removing explicit sensitive-attribute cues does not make the noisy label more reliable. These interactions complicate the assumption that same-group or desensitized demonstrations will improve subgroup outcomes.

We study these interactions in medical in-context prediction, with sex and race as sensitive attributes and prompt-only access to black-box LLMs. Our experiments use CheXpert chest-radiograph data

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.

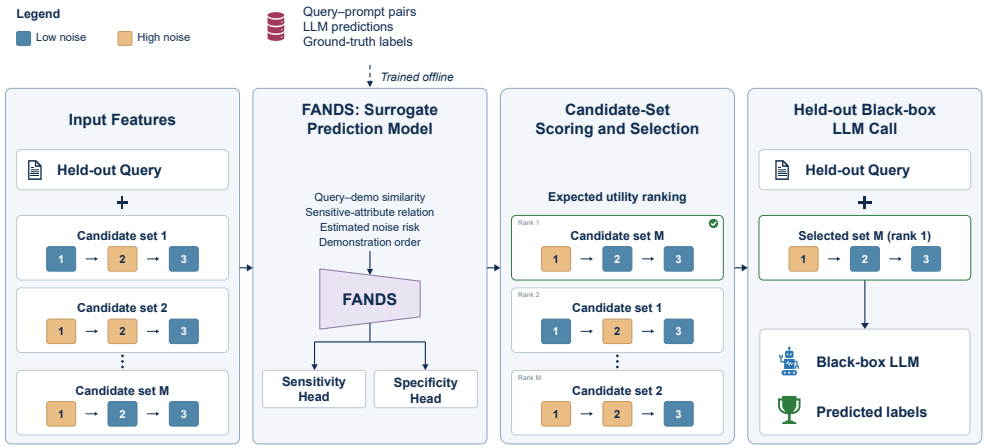

Figure 1: FANDS method overview. The selection surrogate is trained offline using logged query–prompt outcomes from non-test data. At selection time, FANDS represents each eligible prompt configuration using selection-time features, predicts label-conditional sensitivity and specificity, ranks the configurations by expected predictive performance, and sends only the highest-ranked configuration to the held-out black-box LLM.

and Harvard-FairVLMed ophthalmology data. We first use diagnostic regressions to examine how demonstration label noise and sensitive-attribute presentation relate to target-label correctness.

Our analysis shows that label noise changes how fairness-sensitive prompt composition should be interpreted. Demonstrations with incorrect labels are associated with lower target-label correctness across concordant, discordant, and desensitized prompt conditions. The strongest adverse association occurs for an incorrect concordant example, which matches the query's sensitive-attribute group while displaying a noisy label. Desensitization does not eliminate this association between label noise and lower correctness.

These findings motivate *Fairness- and Noise-Aware Demonstration Selection (FANDS)*, our noise-aware demonstration-selection model that considers sensitive-attribute relation, similarity, order, and estimated noise risk together, rather than relying on fixed same-group, different-group, or desensitized retrieval strategies. FANDS improves both overall predictive performance and worst-group performance compared with fixed-retrieval, learned-selection, and fairness-aware baselines. Our contributions are:

1. **We show that demographic concordance does not protect in-context prediction from noisy demonstrations.** Incorrect displayed labels reduce target-label correctness across concordant, discordant, and desensitized prompt conditions.
2. **We find that prompt-composition effects vary across datasets, sensitive attributes, and black-box LLM families.** This motivates query-specific selection rather than fixed same-group, desensitized, or diversity-only retrieval rules.
3. **We introduce FANDS, a prompt-only demonstration selector that combines estimated noise risk with fairness-relevant prompt features.** In completed held-out evaluations, FANDS improves F1 score and balanced accuracy over all the baselines.
4. **We validate FANDS's robustness across datasets, sensitive attributes, black-box LLM families, and demonstration label-noise levels.** These evaluations show that its gains persist under increased noise and extend to lower-performing sensitive-attribute subgroups.

2 Related Work

Demonstration selection under label noise. Few-shot predictions using in-context examples can depend on prompt artifacts, selected examples, and order (Zhao et al., 2021; Lu et al., 2022). Retrieval and learned-selection methods choose demonstrations by similarity or learned relevance, and recent work extends selection to mixture-based retrieval, ordered exemplar optimization, candidate demonstration-pool search, and visual demonstration selection (Liu et al., 2022; Rubin et al., 2022;

Peng et al., 2024; Wang et al., 2024; Wu et al., 2024; Xu et al., 2024). These methods primarily target predictive performance. The effect of demonstration labels varies across tasks: randomizing labels has limited impact in some classification settings (Min et al., 2022), whereas noisy annotations can harm text generation (Gao et al., 2024). Other work identifies reliance on pretrained semantic priors rather than demonstrated input–label relationships (Jang et al., 2024). Medical multimodal evaluations also report sensitivity to irrelevant demonstrations and demonstration order (Rieff et al., 2025). Moreover, noisy examples can be detected, valued, and evaluated under explicit protocol choices (Goswami et al., 2023; Xie et al., 2025).

Fairness-aware prompting and prompt composition. Fairness-aware in-context learning (ICL) establishes that prompt composition can affect subgroup outcomes. Prior work selects or reweights demonstrations to reduce predictive bias, vary demographic representation, or optimize fairness-oriented objectives (Ma et al., 2023; Hu et al., 2024; Aguirre et al., 2024; Halim et al., 2025; Bhaila et al., 2025). However, it is difficult to select demonstrations when sensitive-attribute cues are coupled with label noise or weak clinical relevance. We do not assume that more same-group, minority-group, or desensitized examples are automatically better. Data-selection work outside ICL shows harmful examples can disproportionately affect worst-group performance, even when they are a subset of the available data (Jain et al., 2024).

Label noise and fairness in medical machine learning. Data inclusion, measurement frequency, label definitions, and proxies can change group-conditioned performance and downstream care (Rajkomar et al., 2018; Obermeyer et al., 2019; Vyas et al., 2020). Medical-imaging models can infer race from images even when race is not explicitly provided, so removing sensitive-attribute metadata does not necessarily remove demographic information (Gichoya et al., 2022). In-context medical prediction adds a selection-time version of the same problem because prompt demonstrations expose labels, sensitive-attribute cues, subgroup composition, and similar cases that may be wrong or weakly grounded. Unlike prior work on representation, balancing, or fairness objectives (Hu et al., 2024; Bhaila et al., 2025; Halim et al., 2025), FANDS evaluates fairness-sensitive demonstration selection through subgroup performance.

3 Problem Statement

We study prompt-only in-context prediction for classification. For dataset d and held-out query q, input x_q has true label vector $\mathbf{y}_q \in \{0,1\}^{C_d}$, $y_{q\ell}$ is its ℓth component for $\ell \in \{1,\dots,C_d\}$, C_d is the number of task labels for dataset d, and g_q is the query group for one sensitive attribute analyzed at a time. The queried prediction model is a black-box LLM $\mathcal{M}$, accessed only through prompts. For a candidate demonstration set π, its prediction is

$$\hat{\mathbf{y}}_q(\pi) = \mathcal{M}(x_q;\pi), \tag{1}$$

where $\hat{\mathbf{y}}_q(\pi) \in \{0,1\}^{C_d}$ is the model output for query q under prompt π, and $\hat{y}_{q\ell}(\pi)$ is its ℓth predicted label. The black-box LLM is not fine-tuned, and demonstration selection does not use LLM logits, gradients, hidden states, held-out labels, or held-out LLM responses.

For query q, the selector receives a candidate demonstration pool $\mathcal{D}_q$. Each candidate $i \in \mathcal{D}_q$ consists of displayed content u_i, displayed label vector $\tilde{\mathbf{y}}_i \in \{0,1\}^{C_d}$, displayed label component $\tilde{y}_{i\ell}$, and group attribute g_i. A candidate demonstration set $\pi = (i_1,\dots,i_L)$ contains $L \le K$ distinct demonstrations from $\mathcal{D}_q$, where L is prompt length, K is the maximum number of demonstrations, and later sums over $i \in \pi$ range over these selected demonstration indices. The selector chooses π before the held-out black-box LLM response and held-out label are observed. It may use training labels, group metadata, retrieval features, precomputed estimated noise risk values, and training or validation prompt logs, but not held-out responses or labels. Candidate demonstrations are drawn from non-test pools and are subject-disjoint from the held-out query.

We separate sensitive-attribute composition from demonstration label noise. For a cued demonstration, which explicitly displays the sensitive-attribute value, $r_q(i)$ is $\mathrm{concordant}$ when $g_i = g_q$ and $\mathrm{discordant}$ otherwise. This relation describes only the displayed group relation between candidate i and query q. A desensitized demonstration removes the explicit sensitive-attribute cue while preserving content and the displayed task label. These categories describe only prompt composition. Displayed label quality is treated separately as observed incorrectness in the diagnostic analysis or estimated noise risk at selection time.

Datasets. We use two medical prediction datasets. CheXpert is a chest-radiograph benchmark evaluated as a five-label task (Irvin et al., 2019), and Harvard-FairVLMed is an ophthalmology dataset evaluated here as a binary task (Luo et al., 2024). We analyze sex and race in both datasets, prespecifying female and Black subjects as unprivileged groups and male and White subjects as privileged groups. Further dataset details are in Appendix G.

Selection-time noise-risk estimation. Estimated noise risk measures disagreement between a candidate demonstration's displayed labels and those of nearby eligible examples without querying the black-box LLM. It is lower when displayed labels agree with nearby examples and higher when they disagree. The full details are in Appendix H.1.

Black-box LLMs. We evaluate one GPT-family and one Qwen-family black-box LLM through prompts only. Model identifiers and sizes are in Appendix G.

4 Empirical Analysis: Noise-Aware Demonstration Selection

In this section, we analyze which query and selected demonstrations characteristics are associated with black-box LLM performance in the noisy-label ICL setting. For each evaluated query-prompt pair, we construct one row containing the query case, the selected demonstrations shown to the black-box LLM, and selection-relevant demonstration set features as covariates. We then fit a regression model relating query-label correctness as the dependent outcome to these covariates (e.g., demonstration label noise, sensitive-attribute concordance between the query and demonstrations, desensitization of the sensitive attribute in the demonstrations, query difficulty, query-demonstration similarity). We repeat model training across two datasets, two sensitive attributes (i.e., sex and race), and two black-box LLMs, yielding eight regression models. Empirical-analysis methods are provided in Appendix A, and the complete odds-ratio tables are in Appendix D.

Label noise reverses the value of sensitive-attribute-concordant demonstrations for ICL. Incorrect concordant demonstrations are the most harmful noisy-demonstration condition in our diagnostic analysis. Compared with incorrect discordant and incorrect desensitized alternatives, they produce the steepest probability-scale decline in Figure 2 and have the lowest ORs across all eight diagnostic fits (OR range 0.158–0.671, $p < 0.001$). This risk applies to fairness-aware ICL strategies that select demonstrations by demographic relation or group balance, including within-demographic sampling, same-group selection, and minority-enriched prompting (Ma et al., 2023; Hu et al., 2024; Aguirre et al., 2024; Halim et al., 2025; Bhaila et al., 2025). These strategies can select an incorrect concordant demonstration because it satisfies the demographic criterion while giving the black-box LLM a misleading clinical association. Without noise control, concordance-based selection can turn intended same-group evidence into a source of diagnostic error.

Desensitization is not a label-noise fix or a reliable demonstration-selection strategy. Desensitization does not remove the harm of label noise in noisy-label ICL. Figure 2 shows that incorrect desensitized demonstrations remain below the all-correct-concordant base across all eight settings (OR range 0.208–0.708, $p < 0.001$). It also does not reliably improve selection when labels are correct. Correct desensitized demonstrations have ORs of 0.728–0.981 relative to correct concordant demonstrations, with significant decreases in three of eight settings. Desensitization and demographic-blindness strategies (Rajkomar et al., 2018; Gichoya et al., 2022) do not correct label noise and can discard useful concordant signal, making them unreliable substitutes.

No single demonstration-selection rule generalizes for noisy-label ICL. The regression models do not support a global demonstration-selection rule for noisy labels. Demonstration-count ORs are near null (OR ranges 0.973–1.063), with no significant effects. Unprivileged-query main-effect ORs range from 0.604–1.097 with significant decreases in five settings and no significant increases, and last-incorrect interaction ORs range from 0.866–1.015 with significant decreases only in two of eight settings. These heterogeneous associations support query-specific, fairness- and noise-aware demonstration selection rather than generalized rules.

5 Fairness- and Noise-Aware Demonstration Selection

Motivated by the evidence in Section 4, we propose FANDS, which treats in-context demonstration selection as a prediction problem over candidate demonstration sets. For each held-out query q,

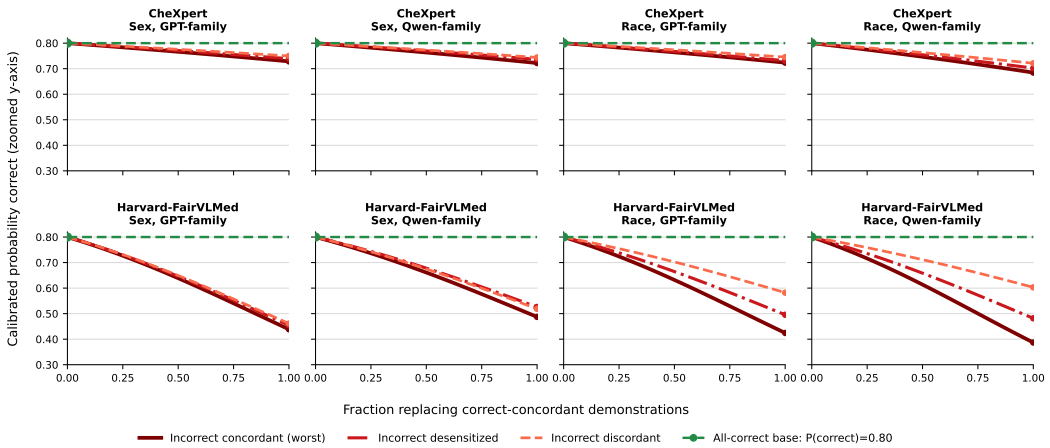

Figure 2: Diagnostic label-noise sensitivity curves. Each curve replaces the correct-concordant reference fraction with the named incorrect component from a common $P(\text{correct}) = 0.80$ base case. Darker red marks incorrect concordant demonstrations, the lowest-odds component in all eight fits.

FANDS represents each candidate demonstration set using only selection-time features (e.g., query-demonstration similarity, sensitive-attribute relation, and estimated noise risk). A learned surrogate model estimates the black-box LLM's expected sensitivity and specificity under each candidate demonstration set, and a scoring rule converts those estimates into the final prompt choice.

The following sections define the feature representation used before the held-out LLM call, the surrogate model that predicts label-conditional LLM behavior, and the scoring rule used to choose the final demonstration set. Figure 1 summarizes this selection pipeline.

Surrogate model development for estimating LLM behavior. The selection surrogate model estimates how the black-box LLM will behave under a candidate demonstration set. For each dataset, sensitive attribute, and black-box LLM combination, the surrogate is trained using only selection-time features and logged black-box LLM outcomes from non-test splits. Section 6 describes how those logged training rows are constructed. This keeps FANDS data-set-agnostic and black-box-LLM-agnostic. Changing the prediction task changes the label dimension and logged outcomes, not the information boundary or the selection rule.

The surrogate has two label-conditional heads. For query q, demonstration set π, and label $\ell \in \{1, \ldots, C_d\}$, the sensitivity head estimates

$$\hat{a}_{q\pi\ell} \approx \Pr(\hat{y}_{q\ell}(\pi) = 1 \mid y_{q\ell} = 1, q, \pi, \ell), \tag{2}$$

and the specificity head estimates

$$\hat{b}_{q\pi\ell} \approx \Pr(\hat{y}_{q\ell}(\pi) = 0 \mid y_{q\ell} = 0, q, \pi, \ell). \tag{3}$$

The sensitivity head is trained on positive-label cases and the specificity head on negative-label cases. For a multi-label task, the heads are evaluated over the task label dimension C_d. For a binary prediction task, the same formulation reduces to the one-label case.

Selection-time features and prior LLM evaluation data. FANDS represents each candidate demonstration set using selection-time surrogate features (e.g., image embeddings, prompt length, query-demonstration similarity, estimated noise risk). The full feature families and excluded held-out or oracle inputs are listed in Appendix H.3. Prior evaluations of the same black-box LLM are needed because they provide the supervised data used to estimate how candidate demonstration sets affect LLM behavior before held-out evaluation.

Demonstration-set scoring and selection. For each query q, candidate demonstration set π, and label ℓ, the surrogate model estimates the black-box LLM's sensitivity and specificity. FANDS ranks candidate demonstration sets by converting these estimates into a scalar score for the trade-off between detecting positive labels and avoiding false positives. A simple illustrative scoring rule is a weighted average of predicted sensitivity and specificity.

$$S_\alpha(q, \pi) = \frac{1}{C_d} \sum_{\ell=1}^{C_d} \left[\alpha \hat{a}_{q\pi\ell} + (1 - \alpha) \hat{b}_{q\pi\ell} \right],$$

where $\hat{a}_{q\pi\ell}$ is predicted sensitivity, $\hat{b}_{q\pi\ell}$ is predicted specificity, C_d is the number of labels, and $\alpha \in [0, 1]$ controls the sensitivity-specificity trade-off.

The full FANDS score combines the surrogate heads with a query-local estimate of label positivity to form expected true-positive, false-negative, true-negative, and false-positive terms. These terms define expected accuracy, expected F1, expected sensitivity, and expected specificity before the held-out black-box LLM response is observed, allowing validation-selected metric trade-offs. The full scoring rule is provided in Appendix H.4. Let $\mathcal{P}_q$ denote the candidate prompt pool, containing all eligible prompt configurations formed from $\mathcal{D}_q$ using one, two, or three demonstrations in cued or desensitized form. FANDS scores every configuration in $\mathcal{P}_q$ and selects the highest-scoring one before querying the black-box LLM.

$$\pi^\star(q) = \arg\max_{\pi \in \mathcal{P}_q} S(q, \pi). \tag{4}$$

where $\pi^\star(q)$ is the selected demonstration set used for the final black-box LLM call.

6 Experiments

Data setup. We evaluate CheXpert (five-label chest radiographs) and Harvard-FairVLMed (binary glaucoma prediction). For each dataset, sensitive attribute, and LLM, subject-disjoint splits contain 700 query subjects for surrogate training, 200 for validation, and 100 for held-out evaluation. The surrogate-training split yields 70,000 logged query-prompt evaluations for each analysis setting. Surrogates are trained separately for each combination, with no sharing of logged outcomes across datasets, attributes, or black-box LLMs. We use validation outcomes and validation LLM calls only for method selection, and the held-out split only for final evaluation. All methods use the same query-specific, subject-disjoint candidate demonstration pool drawn from non-test data. Details on labels, attributes, splits, candidate-pool construction, and prompts are in Appendix G.

Metrics. Primary metrics are macro F1 and macro balanced accuracy for CheXpert, and F1 and balanced accuracy for Harvard-FairVLMed. For subgroup evaluation, we report worst-group F1, i.e., the lower F1 across the two sensitive-attribute groups. Appendix I includes detailed definitions.

Ablations. Subgroup analysis. We report selected demonstration set performance separately for female and male subjects in sex analyses and Black and White subjects in race analyses, and summarize performance for the lower-performing group using worst-group F1. **Component-ablation analysis.** We retrain five component-ablation variants in every dataset, sensitive-attribute, and black-box LLM setting. Four variants remove the estimated noise-risk, order-sensitive, similarity/redundancy, or group-relation feature block; the fifth replaces the label-conditional sensitivity and specificity heads with a single scalar utility surrogate. Appendix L defines these components and provides the complete protocol. **Noise-stress analysis.** We perturb displayed demonstration labels in the candidate pools while keeping query labels unchanged. We compare FANDS with the baselines under asymmetric group-noise injection, which affects only unprivileged demonstrations, and pool-wide injection, which affects both groups. The target row-weighted mean estimated noise risk within each affected group is 40% for CheXpert and 50% for Harvard-FairVLMed. Details are provided in Appendix E.

Baselines. Each baseline selects three cued demonstrations from the same query-specific candidate pool before the held-out LLM response is observed. The baselines are as follows. **Full-pool relevance-risk retrieval** ranks all eligible prompts by a fixed score that rewards average query-demonstration similarity and penalizes estimated noise risk (Liu et al., 2022; Rubin et al., 2022; Gao et al., 2024). **Same-group relevance-risk retrieval** applies the same score after restricting candidates to the query's sensitive-attribute group (Ma et al., 2023; Hu et al., 2024; Aguirre et al., 2024; Bhaila et al., 2025), and **MMR diversity retrieval** greedily fills the three slots with maximal marginal relevance, favoring query similarity while penalizing redundancy (Carbonell and Goldstein, 1998). **Visual EPR** learns a visual query-demonstration retriever from logged black-box LLM outcomes (Rubin et al., 2022). **Multimodal EASE** learns an order-sensitive reward model over three-demonstration sequences (Wu et al., 2024). **FCG fairness-aware selection** searches group- and label-stratified candidates using validation-selected utility, fairness, diversity, and relevance criteria (Hu et al., 2024). Complete scoring, implementation, and tuning details are provided in Appendix H.2.

Table 1: Overall and subgroup selected-prompt F1 score. CheXpert reports macro F1; Harvard-FairVLMed reports binary F1. In each LLM family, the first column reports overall F1. Female and Black are unprivileged groups; Male and White are privileged groups. Values are point estimates ± symmetric 95% subject-bootstrap margins. Bold identifies the best unrounded value within each dataset, attribute, LLM, and evaluation column.

Dataset	Attribute	Policy	GPT-family			Qwen-family		
			F1 score ↑	Unprivileged group (F1 ↑)	Privileged group (F1 ↑)	F1 score ↑	Unprivileged group (F1 ↑)	Privileged group (F1 ↑)
CheXpert	Sex	FANDS (ours)	**.68 ± .06**	**.69 ± .08**	**.66 ± .10**	**.66 ± .07**	**.63 ± .10**	**.69 ± .09**
		Full-pool relevance-risk	.60 ± .08	.61 ± .10	.59 ± .13	.61 ± .08	.59 ± .11	.63 ± .12
		Same-group relevance-risk	.61 ± .08	.61 ± .11	.61 ± .11	.58 ± .09	.56 ± .13	.59 ± .13
		MMR diversity	.60 ± .07	.61 ± .09	.59 ± .11	.56 ± .08	.55 ± .11	.58 ± .12
		Visual EPR	.63 ± .06	.62 ± .09	.64 ± .09	.64 ± .07	.62 ± .09	.66 ± .10
		Multimodal EASE	.60 ± .08	.60 ± .11	.59 ± .13	.52 ± .09	.50 ± .13	.53 ± .13
		FCG fairness-aware selection	.58 ± .07	.59 ± .09	.57 ± .11	.59 ± .09	.58 ± .12	.60 ± .13
	Race	FANDS (ours)	**.63 ± .07**	**.64 ± .11**	**.63 ± .10**	**.67 ± .07**	**.64 ± .10**	**.70 ± .09**
		Full-pool relevance-risk	.57 ± .08	.55 ± .12	.58 ± .10	.55 ± .09	.48 ± .14	.60 ± .11
		Same-group relevance-risk	.51 ± .08	.49 ± .12	.52 ± .10	.49 ± .09	.42 ± .13	.56 ± .13
		MMR diversity	.60 ± .07	.60 ± .10	.59 ± .10	.54 ± .09	.50 ± .12	.57 ± .13
		Visual EPR	.59 ± .07	.58 ± .10	.59 ± .09	.50 ± .09	.47 ± .12	.53 ± .12
		Multimodal EASE	.54 ± .07	.56 ± .12	.51 ± .09	.46 ± .08	.38 ± .12	.52 ± .11
		FCG fairness-aware selection	.56 ± .07	.57 ± .10	.54 ± .11	.50 ± .09	.42 ± .13	.57 ± .13
Harvard-FairVLMed	Sex	FANDS (ours)	**.78 ± .09**	**.75 ± .14**	**.82 ± .12**	**.70 ± .11**	**.71 ± .15**	.68 ± .17
		Full-pool relevance-risk	.73 ± .10	.64 ± .17	**.82 ± .12**	.65 ± .11	.60 ± .18	.70 ± .16
		Same-group relevance-risk	.74 ± .10	.71 ± .15	.76 ± .14	.63 ± .12	.63 ± .17	.62 ± .17
		MMR diversity	.66 ± .11	.64 ± .16	.68 ± .15	.63 ± .12	.63 ± .17	.62 ± .17
		Visual EPR	.08 ± .09	.08 ± .13	.08 ± .13	.04 ± .07	.00 ± .00	.08 ± .13
		Multimodal EASE	.65 ± .09	.66 ± .13	.65 ± .12	.70 ± .09	.70 ± .13	**.70 ± .13**
		FCG fairness-aware selection	.53 ± .11	.51 ± .16	.56 ± .16	.50 ± .13	.49 ± .19	.51 ± .19
	Race	FANDS (ours)	**.76 ± .09**	**.73 ± .14**	**.79 ± .13**	.63 ± .12	.63 ± .18	.63 ± .16
		Full-pool relevance-risk	.73 ± .10	.68 ± .17	.78 ± .13	**.69 ± .11**	.62 ± .18	**.75 ± .14**
		Same-group relevance-risk	.72 ± .10	.67 ± .16	.78 ± .13	.67 ± .11	**.70 ± .16**	.64 ± .17
		MMR diversity	.65 ± .11	.63 ± .17	.67 ± .15	.58 ± .12	.65 ± .17	.51 ± .19
		Visual EPR	.21 ± .14	.21 ± .21	.21 ± .21	.08 ± .09	.08 ± .13	.08 ± .13
		Multimodal EASE	.62 ± .11	.67 ± .15	.57 ± .17	.31 ± .15	.39 ± .23	.21 ± .21
		FCG fairness-aware selection	.63 ± .11	.63 ± .16	.63 ± .14	.57 ± .12	.69 ± .15	.40 ± .19

7 Experimental Results

In this section, we address three research questions. First, can fairness- and noise-aware demonstration selection (FANDS) improve predictive performance over fixed retrieval, learned selection, and fairness-aware baselines? Second, does FANDS improve worst-group performance across sensitive-attribute subgroups? Finally, do selected-demonstration gains hold as demonstration label noise increases?

7.1 FANDS improves demonstration selection for predictive performance

FANDS achieves the highest overall F1 in all four CheXpert settings and three of the four Harvard-FairVLMed settings compared with the baselines. The exception is Harvard-FairVLMed race analysis under Qwen-family instead favors full-pool relevance-risk retrieval (Table 1). Harvard-FairVLMed uses highly similar demonstrations (Appendix N), which may leave less room for FANDS to improve the demonstration selection. Balanced-accuracy results are in Table 2.

Component contributions. To identify which components support these gains, we compare complete FANDS with five ablated variants (Figure 3). Retaining every component improves mean F1 across the four GPT-family settings, with order-sensitive features making the largest mean contribution. Complete FANDS achieves the highest or tied-highest F1 in all four settings, supporting the combined use of these components. Appendix L provides the protocol and balanced-accuracy results.

Sample efficiency. We examine FANDS efficiency and dependency on the training budget. Reducing the budget from 70,000 to 30,000 query–prompt outcomes per setting retains 95.4% of mean F1 and 96.5% of mean balanced accuracy across the eight settings, while preserving positive F1 margins over the baselines. Thus, FANDS retains much of its predictive performance with 57.1% fewer training outcomes, although the complete budget provides a measurable benefit (Appendix M).

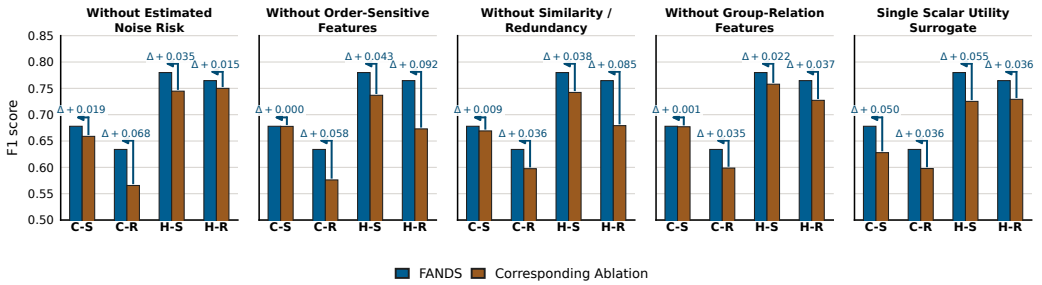

Figure 3: GPT-family component ablation on held-out F1. Each panel compares FANDS with the ablation named in its heading. Arrows report $\Delta = \text{FANDS} - \text{ablation}$. C-S denotes CheXpert sex, C-R CheXpert race, H-S Harvard-FairVLMed sex, and H-R Harvard-FairVLMed race. CheXpert reports macro F1 and Harvard-FairVLMed reports binary F1.

Candidate-similarity sensitivity. To assess whether FANDS performance depend on the original candidate pools, we vary query–demonstration similarity within CheXpert. FANDS retains higher equal-setting mean F1 and balanced accuracy than the strongest of baselines at every tested level. These results support robustness of its aggregate advantage across the tested candidate-similarity range (Appendix N).

7.2 Persistence of gains across population subgroups

We next assess whether FANDS's overall gains extend across sensitive-attribute subgroups. FANDS attains or ties the highest F1 in 13 of the 16 subgroup evaluations and does so for both groups in six of eight settings (Table 1). It achieves the highest worst-group F1 in six of eight settings against all evaluated baselines (Appendix C). These results indicate that the gains generally extend to the lower-performing subgroup rather than being concentrated in the better-performing group. Corresponding subgroup balanced-accuracy results appear in Appendix B.

To assess whether the noise-risk estimator feature systematically assigns greater risk to unprivileged groups, we deliberately invert displayed demonstration labels and compare estimated noise risk $\widehat{\epsilon}_i$ across sensitive-attribute groups. Estimated risk is not consistently higher for unprivileged groups, and the noise-risk estimator detects risk generally similar across groups. FANDS also does not exclude demonstrations solely because of estimated risk, as unprivileged demonstrations constitute 40.3–61.2% of selected demonstration slots (Appendix K.1).

7.3 Selected-demonstration gains under increased demonstration label noise

Label noise is a key robustness concern for demonstration selection because noisy demonstrations can propagate misleading contextual signals into black-box LLM prompts, especially when noise is uneven across population subgroups. We therefore tested whether the selected-demonstration advantage persists under two CheXpert 40% target-risk settings and the corresponding 50% settings for Harvard-FairVLMed, with the stress protocol and dataset context summarized in Appendix E.

Figure 4 reports the resulting F1-family scores. In panel A, FANDS remains above the condition-selected best baseline in all available stress comparisons. These CheXpert point estimates suggest that selected-demonstration gains can persist under increased label noise. Absolute performance declines in most settings (FANDS macro F1 decreases in six of eight comparisons), while the best baseline decreases in all eight. The corresponding macro balanced-accuracy results in Appendix F show a similar degradation pattern under the same exploratory boundary. In panel B, FANDS leads the condition-selected best baseline in five of eight stress comparisons; the race analysis under Qwen-family and the pool-wide sex analysis under Qwen-family favor the best baseline. Because the Harvard-FairVLMed candidate pools are much more visually similar than the CheXpert pools, this stress test separates visual relevance from displayed-label reliability.

We next test whether FANDS depends on precise estimated noise-risk values, we add random error of up to 0.20 to noise-risk scores $\widehat{\epsilon}_i$ and repeat selection without retraining. Across perturbation levels, mean F1 and balanced accuracy remain close to their unperturbed values, and FANDS retains higher

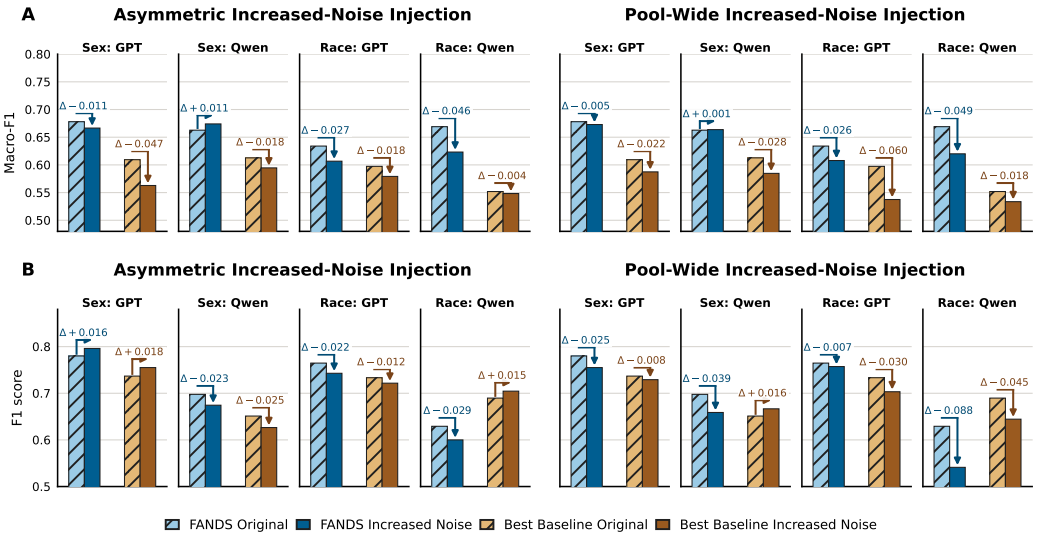

Figure 4: Label-noise ablation on F1-family metrics. Panel A shows CheXpert 40% target-risk macro F1; panel B shows Harvard-FairVLMed 50% target-risk F1 score. The left panels use asymmetric group-noise injection; the right panels use pool-wide noise injection. Hatched bars are original values, solid bars are post-injection values, and arrows show deltas. The best baseline is selected within condition by the plotted F1-family metric.

F1 than the strongest baseline in seven of eight settings at every level (Appendix K.2). These results support robustness to the tested errors in estimated noise risk.

In several race analyses, asymmetric subgroup noise produces degradation comparable to overall noise injection, so subgroup-concentrated label corruption can affect retrieval quality even when the privileged-group pool is unchanged. This is consistent with evidence that medical-imaging label noise and label bias can differ across demographic subgroups (Zhang et al., 2022).

8 Discussion and Conclusion

Across the diagnostic analyses, demonstration label noise is consistently associated with lower target-label accuracy in the CheXpert and Harvard-FairVLMed analyses. FANDS responds by combining estimated noise risk, group relation, ordering, and similarity rather than treating same-group, different-group, or desensitized retrieval as a fairness fix. The diagnostic regressions are coefficient-scale evidence, so they motivate selected-prompt evaluation rather than establishing final subgroup accuracy or fairness mitigation.

Our results support gains in F1 and balanced accuracy over baselines in most evaluated settings. Improvements in worst-group F1 show that these gains often extend to the lower-performing subgroup, without establishing subgroup parity. The two datasets also differ in label-noise evidence and visual similarity. CheXpert includes explicit uncertainty annotations and less visually similar demonstrations, while Harvard-FairVLMed depends on estimated noise risk and has highly similar demonstrations. Candidate similarity affects the relevance of the available in-context evidence, but does not alone determine whether FANDS improves over fixed retrieval. Together, these differences make the evaluation a useful generalizability check across image prediction tasks and demonstration pools.

The intended benefit is more careful demonstration selection when subgroup performance matters, labels may be noisy, and black-box LLM queries are limited. Key deployment considerations are generalization beyond the evaluated medical benchmarks, treating estimated noise risk as a proxy rather than a ground-truth label-quality signal, and avoiding deployment claims from descriptive subgroup evidence. Deployment-oriented use would require privacy review, local validation, paired subgroup testing, clinical oversight, and prospective monitoring.

AI USE STATEMENT

Generative AI tools were used to assist with experimental implementation and with drafting, reviewing, and polishing the manuscript. The authors reviewed and verified all AI-assisted code, experimental outputs, and text. The authors take full responsibility for the final content of this work, including its analyses, claims, and artifacts.

REFERENCES

[1] Carlos Aguirre, Kuleen Sasse, Isabel Cachola, and Mark Dredze. Selecting Shots for Demographic Fairness in Few-Shot Learning with Large Language Models. In Daryna Dementieva, Oana Ignat, Zhijing Jin, Rada Mihalcea, Giorgio Piatti, Joel Tetreault, Steven Wilson, and Jieyu Zhao, editors, *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 50–67, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4pi-1.4.

[2] Karuna Bhaila, Minh-Hao Van, Kennedy Edemacu, Chen Zhao, Feng Chen, and Xintao Wu. Fair In-Context Learning via Latent Concept Variables. In *2025 IEEE International Conference on Big Data (BigData)*, pages 1827–1836, 2025. doi: 10.1109/BigData66926.2025.11402286.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[4] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, August 1998. Association for Computing Machinery. ISBN 978-1-58113-015-7. doi: 10.1145/290941.291025.

[5] Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. On the Noise Robustness of In-Context Learning for Text Generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 16569–16600. Curran Associates, Inc., 2024. doi: 10.52202/079017-0528.

[6] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, June 2022. ISSN 2589-7500. doi: 10.1016/S2589-7500(22)00063-2.

[7] Mononito Goswami, Vedant Sanil, Arjun Choudhry, Arvind Srinivasan, Chalisa Udompanyawit, and Artur Dubrawski. AQuA: A Benchmarking Tool for Label Quality Assessment. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 79792–79807. Curran Associates, Inc., 2023. doi: 10.52202/075280-3494.

[8] Sadaf Md Halim, Chen Zhao, Xintao Wu, Latifur Khan, Christan Grant, Fariha Ishrat Rahman, and Feng Chen. Let The Jury Decide: Fair Demonstration Selection for In-Context Learning through Incremental Greedy Evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18914–18931, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.968.

[9] Jingyu Hu, Weiru Liu, and Mengnan Du. Strategic Demonstration Selection for Improved Fairness in LLM In-Context Learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7460–7475, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.425.

[10] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.3301590.

[11] Saachi Jain, Kimia Hamidieh, Kristian Georgiev, Andrew Ilyas, Marzyeh Ghassemi, and Aleksander Mądry. Improving Subgroup Robustness via Data Selection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 94490–94511. Curran Associates, Inc., 2024. doi: 10.52202/079017-2995.

[12] Joonwon Jang, Sanghwan Jang, Wonbin Kweon, Minjin Jeon, and Hwanjo Yu. Rectifying Demonstration Shortcut in In-Context Learning. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4294–4321, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.242.

[13] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What Makes Good In-Context Examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10.

[14] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556.

[15] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, Yi Fang, and Mengyu Wang. FairCLIP: Harnessing Fairness in Vision-Language Learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12289–12301, 2024. doi: 10.1109/CVPR52733.2024.01168.

[16] Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. Fairness-guided Few-shot Prompting for Large Language Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 43136–43155. Curran Associates, Inc., 2023. doi: 10.52202/075280-1870.

[17] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759.

[18] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. doi: 10.1126/science.aax2342.

[19] Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting Demonstration Selection Strategies in In-Context Learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.492.

[20] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00965-w.

[21] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, 169(12):866–872, December 2018. ISSN 0003-4819. doi: 10.7326/M18-1990.

[22] Melanie Rieff, Maya Varma, Ossian Rabow, Subathra Adithan, Julie Kim, Ken Chang, Hannah Lee, Nidhi Rohatgi, Christian Bluethgen, Mohamed Muneer, Jean-Benoit Delbrouck, and Michael Moor. SMMILE: An expert-driven benchmark for multimodal medical in-context learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2180.

[23] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning To Retrieve Prompts for In-Context Learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.191.

[24] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *Biocomputing 2021*, pages 232–243. WORLD SCIENTIFIC, October 2020. ISBN 978-981-12-3269-5. doi: 10.1142/9789811232701_0022.

[25] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, 383(9):874–882, August 2020. ISSN 0028-4793. doi: 10.1056/NEJMms2004740.

[26] Song Wang, Zihan Chen, Chengshuai Shi, Cong Shen, and Jundong Li. Mixture of Demonstrations for In-Context Learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 88091–88116. Curran Associates, Inc., 2024. doi: 10.52202/079017-2796.

[27] Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, Wenyang Hu, Yao Shu, See-Kiong Ng, Patrick Jaillet, and Bryan Kian Hsiang Low. Prompt Optimization with EASE? Efficient Ordering-aware Automated Selection of Exemplars. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 122706–122740. Curran Associates, Inc., 2024. doi: 10.52202/079017-3899.

[28] Shan Xie, Man Luo, Chadly Daniel Stern, Mengnan Du, and Cheng Lu. DemoShapley: Valuation of Demonstrations for In-Context Learning. In *2025 IEEE International Conference on Big Data (BigData)*, pages 4081–4090, 2025. doi: 10.1109/BigData66926.2025.11402298.

[29] Chengming Xu, Chen Liu, Yikai Wang, Yuan Yao, and Yanwei Fu. Towards Global Optimal Visual In-Context Learning Prompt Selection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 74945–74965. Curran Associates, Inc., 2024. doi: 10.52202/079017-2384.

[30] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. Improving the Fairness of Chest X-ray Classifiers. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 204–233. PMLR, April 2022.

[31] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-shot Performance of Language Models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, July 2021.

APPENDIX OVERVIEW

CONTENTS

Appendix	Title	Page
A	Empirical Analysis Methods	15
B	Overall and Subgroup Balanced Accuracy	16
C	Worst-Group F1	17
D	Full Diagnostic Odds-Ratio Tables	18
E	Exploratory Noise-Stress Ablation Protocol	20
F	Label-Noise Injection Balanced Accuracy Figures	21
G	Data Processing	21
H	Method Details	22
I	Metric Definitions and Statistical Uncertainty	26
J	FANDS Selection Surrogate Architecture and Training	27
K	Estimated Noise-Risk Audits	28
L	Component-Level Ablations	30
M	Sample Efficiency	31
N	Candidate-Similarity Sensitivity	32

A Empirical Analysis Methods

This appendix records the method details behind the empirical analysis in Section 4. Section 3 defines the black-box LLM prediction task and candidate demonstration pool. This appendix only records how we construct the diagnostic regression rows. Here q indexes a query, i indexes a candidate demonstration, π is a candidate demonstration set, g_q and g_i are the analyzed sensitive-attribute groups, and C_d is the number of task labels for dataset d. Result-specific notation introduced here is local to the diagnostic row, probability-curve, and table calculations.

Controlled diagnostic query-prompt construction. The diagnostic regression is trained on a controlled panel of evaluated query-prompt rows rather than on only the prompts that a selector would choose. For each analyzed query and sensitive attribute, we construct prompts with $L \in \{1, 2, 3\}$ demonstrations so that the rows span four planned axes. The axes are query group, demonstration group relation, displayed-label quality, and prompt position. Cued demonstrations expose the sensitive-attribute value in the demonstration text. Desensitized variants remove that explicit cue while preserving the content and displayed task label.

We use controlled prompt variants to separate demographic relation from displayed-label correctness. For a cued candidate demonstration i, the group relation to query q is defined from g_q, the query group for the analyzed sensitive attribute, and g_i, the candidate demonstration group.

$$r_q(i) = \begin{cases} \text{concordant}, & g_i = g_q, \\ \text{discordant}, & g_i \neq g_q. \end{cases}$$

To create the label-noise conditions used by the regression, we hold the query, the demonstration content, and prompt position fixed while changing the displayed demonstration label into a correct or incorrect version for the evaluated task. This produces correct-concordant, incorrect-concordant, correct-discordant, incorrect-discordant, correct-desensitized, and incorrect-desensitized prompt components. These manipulated labels are diagnostic interventions used to study prompt-composition sensitivity. They are not available to the selector as oracle demonstration-correctness features at selection time.

Grouped-binomial diagnostic model. The diagnostic regression asks whether displayed-label quality and group relation are associated with target-label correctness after accounting for prompt length, query difficulty, query-demonstration similarity, and repeated evaluations of the same query. For query-prompt row j, let $q[j]$ be the corresponding query, Y_j the number of correctly predicted target labels, and N_j the number of evaluated target labels. Let $\pi[j]$ be the corresponding prompt. For each analysis setting, we fit

$$Y_j \sim \text{Binomial}(N_j, \theta_j), \tag{5}$$

$$\text{logit}(\theta_j) = X_j^\top \beta + u_{q[j]}, \tag{6}$$

$$u_{q[j]} \sim \mathcal{N}(0, \sigma_u^2). \tag{7}$$

Here θ_j is the fitted probability that an evaluated target label is predicted correctly for row j, X_j is the fixed-effect design row, β is the fixed-effect coefficient vector, and $u_{q[j]}$ is a query-level random intercept. The random intercept accounts for multiple prompt variants evaluated on the same query.

The fixed-effect design separates prompt quantity, prompt composition, query context, and selected interactions. Let $\mathcal{T}$ contain the five non-reference prompt-composition components. These are correct discordant, incorrect concordant, incorrect discordant, correct desensitized, and incorrect desensitized. Correct concordant demonstrations are the omitted composition reference. The fitted linear predictor is

$$\begin{aligned} X_j^\top \beta = \beta_0 &+ \beta_D \text{diff}_{q[j]} + \beta_U U_{q[j]} \\ &+ \delta_2 \mathbb{1}[L_j = 2] + \delta_3 \mathbb{1}[L_j = 3] + \sum_{t \in \mathcal{T}} \beta_t p_{jt} \\ &+ \beta_{\text{last}} I_j^{\text{last}-\text{inc}} + \beta_s \bar{s}_j + \beta_{\text{red}} \text{red}_j \\ &+ \sum_{t \in \mathcal{T}} \gamma_t U_{q[j]} p_{jt} + \gamma_{\text{last}} U_{q[j]} I_j^{\text{last}-\text{inc}}. \end{aligned}$$

In this expression, $\mathrm{diff}_{q[j]}$ is a task-difficulty covariate, $U_{q[j]}$ is one for the prespecified unprivileged query group and zero for the privileged or reference group, L_j is prompt length, p_{jt} is the fraction of filled demonstration slots in component t, $I_j^{\mathrm{last-inc}}$ indicates that the final demonstration is incorrect, $\bar{s}_j = \bar{s}_{q[j],\pi[j]}$ is mean query-demonstration similarity, and red_j is the prompt-redundancy covariate when the analysis setting supports it. β_0 is the intercept, δ_2 and δ_3 are prompt-length coefficients, the β terms are main-effect coefficients, and the γ terms test whether prompt-composition associations differ by query group. Coefficients are reported as odds ratios and interpreted as diagnostic associations rather than causal effects.

Figure 2 calculation. The probability curves in Figure 2 convert the prompt-composition odds ratios into a common probability scale. Let $R = 1$ denote a correct target-label prediction, and let $\omega \in [0, 1]$ be the replacement fraction assigned to one plotted component c. For each curve, the plotted replacement path sets $p_c(\omega) = \omega$, sets the omitted correct-concordant fraction to $p_{\mathrm{cc}}(\omega) = 1 - \omega$, and sets the other prompt-composition fractions to zero. Because correct concordant is the reference component, its coefficient is absorbed into the base case. With the common base logit $\mu_0 = \mathrm{logit}(0.80)$, the plotted probability is

$$\Pr(R = 1 \mid \omega, c) = \mathrm{logit}^{-1}(\mu_0 + \beta_c \omega)\,,$$

where β_c is the fitted log-odds coefficient for replacing the correct-concordant fraction with component c. The corresponding odds ratio is $\exp(\beta_c)$ for the full $\omega = 0$ to $\omega = 1$ replacement. Thus the curve is a single compositional contrast. The named component increases by the same fraction that the correct-concordant reference decreases.

B Overall and Subgroup Balanced Accuracy

Table 2 complements the F1 results in Table 1 by reporting balanced accuracy overall and within each sensitive-attribute group.

Table 2: Overall and subgroup selected-prompt balanced accuracy. CheXpert reports macro balanced accuracy; Harvard-FairVLMed reports binary balanced accuracy. Within each LLM family, the first column reports overall balanced accuracy. Female and Black are unprivileged groups; Male and White are privileged groups. Values are point estimates ± symmetric 95% subject-bootstrap margins. Higher values are better; bold identifies the best unrounded value within each dataset, attribute, LLM, and evaluation column.

Dataset	Attribute	Policy	GPT-family			Qwen-family		
			Balanced accuracy ↑	Unprivileged group	Privileged group	Balanced accuracy ↑	Unprivileged group	Privileged group
CheXpert	Sex	FANDS (ours)	**.73 ± .04**	**.74 ± .06**	**.72 ± .07**	**.70 ± .05**	**.69 ± .07**	**.71 ± .07**
		Full-pool relevance-risk	.69 ± .05	.69 ± .06	.69 ± .07	.70 ± .05	.68 ± .06	.71 ± .07
		Same-group relevance-risk	.70 ± .04	.69 ± .06	.70 ± .06	.67 ± .05	.65 ± .07	.69 ± .07
		MMR diversity	.68 ± .04	.69 ± .06	.68 ± .06	.67 ± .04	.65 ± .06	.69 ± .06
		Visual EPR	.70 ± .04	.70 ± .06	.70 ± .06	.65 ± .05	.63 ± .08	.67 ± .08
		Multimodal EASE	.69 ± .05	.68 ± .07	.70 ± .07	.66 ± .04	.65 ± .06	.67 ± .06
		FCG fairness-aware selection	.68 ± .04	.68 ± .05	.67 ± .06	.67 ± .05	.67 ± .06	.68 ± .07
	Race	FANDS (ours)	**.71 ± .05**	**.72 ± .07**	**.69 ± .06**	**.70 ± .05**	**.67 ± .07**	**.73 ± .07**
		Full-pool relevance-risk	.68 ± .04	.68 ± .06	.68 ± .06	.66 ± .05	.64 ± .06	.68 ± .07
		Same-group relevance-risk	.64 ± .04	.64 ± .06	.64 ± .06	.64 ± .04	.62 ± .06	.66 ± .06
		MMR diversity	.69 ± .04	.70 ± .06	.68 ± .06	.65 ± .05	.64 ± .06	.66 ± .07
		Visual EPR	.67 ± .04	.67 ± .07	.66 ± .06	.63 ± .05	.62 ± .07	.63 ± .06
		Multimodal EASE	.65 ± .04	.67 ± .07	.63 ± .06	.62 ± .04	.59 ± .05	.64 ± .06
		FCG fairness-aware selection	.67 ± .04	.68 ± .06	.65 ± .07	.63 ± .05	.61 ± .06	.66 ± .07
Harvard-FairVLMed	Sex	FANDS (ours)	**.78 ± .08**	**.74 ± .12**	**.82 ± .10**	**.74 ± .08**	**.74 ± .12**	**.74 ± .11**
		Full-pool relevance-risk	.75 ± .08	.68 ± .13	**.82 ± .11**	.70 ± .08	.68 ± .12	.72 ± .12
		Same-group relevance-risk	.75 ± .09	**.74 ± .12**	.76 ± .12	.69 ± .08	.70 ± .12	.68 ± .12
		MMR diversity	.66 ± .09	.64 ± .14	.68 ± .13	.68 ± .09	.70 ± .12	.66 ± .13
		Visual EPR	.52 ± .03	.52 ± .04	.52 ± .04	.51 ± .02	.50 ± .00	.52 ± .04
		Multimodal EASE	.49 ± .03	.50 ± .06	.48 ± .04	.60 ± .07	.60 ± .10	.60 ± .10
		FCG fairness-aware selection	.49 ± .10	.46 ± .14	.52 ± .14	.58 ± .09	.58 ± .13	.58 ± .13
	Race	FANDS (ours)	**.76 ± .08**	**.74 ± .12**	**.78 ± .11**	.67 ± .09	.70 ± .12	.64 ± .14
		Full-pool relevance-risk	**.76 ± .08**	**.74 ± .11**	**.78 ± .12**	**.73 ± .08**	.70 ± .11	**.76 ± .12**
		Same-group relevance-risk	.74 ± .08	.70 ± .12	**.78 ± .12**	.71 ± .09	**.74 ± .12**	.68 ± .12
		MMR diversity	.64 ± .09	.64 ± .14	.64 ± .13	.64 ± .09	.70 ± .12	.58 ± .13
		Visual EPR	.54 ± .05	.54 ± .08	.54 ± .08	.52 ± .03	.52 ± .04	.52 ± .04
		Multimodal EASE	.62 ± .10	.66 ± .14	.58 ± .14	.59 ± .05	.62 ± .08	.56 ± .06
		FCG fairness-aware selection	.59 ± .09	.62 ± .14	.56 ± .13	.60 ± .09	.68 ± .13	.52 ± .13

C WORST-GROUP F1

Table 3 reports worst-group F1 for every selection policy. FANDS obtains the highest worst-group F1 among all the evaluated selection baseline in six of the eight settings.

Table 3: Worst-group selected-prompt F1 score. Values are point estimates ± symmetric 95% subject-bootstrap margins. CheXpert reports macro F1 across findings, whereas Harvard-FairVLMed reports binary F1. Higher values are better; bold marks the best value within each setting.

Dataset	Attribute	Policy	GPT-family	Qwen-family
CheXpert	Sex	FANDS (ours)	**.66 ± .09**	**.63 ± .10**
		Full-pool relevance-risk	.59 ± .13	.59 ± .11
		Same-group relevance-risk	.61 ± .12	.56 ± .13
		MMR diversity	.59 ± .10	.55 ± .12
		Visual EPR	.62 ± .10	.62 ± .09
		Multimodal EASE	.59 ± .13	.50 ± .14
		FCG fairness-aware selection	.57 ± .11	.58 ± .14
	Race	FANDS (ours)	**.63 ± .11**	**.64 ± .09**
		Full-pool relevance-risk	.55 ± .11	.48 ± .14
		Same-group relevance-risk	.49 ± .12	.42 ± .13
		MMR diversity	.59 ± .10	.50 ± .11
		Visual EPR	.58 ± .10	.47 ± .12
		Multimodal EASE	.51 ± .10	.38 ± .12
		FCG fairness-aware selection	.54 ± .11	.42 ± .13
Harvard-FairVLMed	Sex	FANDS (ours)	**.75 ± .13**	.68 ± .17
		Full-pool relevance-risk	.64 ± .16	.60 ± .17
		Same-group relevance-risk	.71 ± .15	.62 ± .19
		MMR diversity	.64 ± .15	.62 ± .18
		Visual EPR	.08 ± .08	.00 ± .00
		Multimodal EASE	.65 ± .13	**.70 ± .14**
		FCG fairness-aware selection	.51 ± .16	.49 ± .19
	Race	FANDS (ours)	**.73 ± .13**	.63 ± .19
		Full-pool relevance-risk	.68 ± .16	.62 ± .17
		Same-group relevance-risk	.67 ± .15	**.64 ± .16**
		MMR diversity	.63 ± .16	.51 ± .17
		Visual EPR	.21 ± .21	.08 ± .08
		Multimodal EASE	.57 ± .16	.21 ± .21
		FCG fairness-aware selection	.63 ± .17	.40 ± .19

D Full Diagnostic Odds-Ratio Tables

Tables 4 and 5 report the full grouped-binomial fixed-effect odds ratios, including the prompt-composition rows used for Figure 2. Reference levels are shown in the table body. Stars mark the p-value thresholds stated in each caption.

Table 4: Full CheXpert fixed-effect odds ratios. Reference levels are shown in the table. Prompt-composition ORs replace the correct-concordant fraction. Stars mark * $p < .05$, ** $p < .001$.

Term	Sex GPT-family OR	Sex Qwen-family OR	Race GPT-family OR	Race Qwen-family OR
Core covariates				
Intercept	31.138**	21.376**	22.028**	13.436**
Task difficulty, per 1-unit increase	0.006**	0.010**	0.010**	0.017**
Query group				
Privileged query	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Unprivileged query	0.719**	0.604**	0.841**	0.991
Prompt count				
One demonstration	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Two demonstrations	0.976	1.048	0.994	1.020
Three demonstrations	0.995	1.040	0.999	1.032
Prompt-composition main effects				
Correct-concordant fraction	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Correct-discordant fraction	0.992	0.927*	0.885**	0.998
Incorrect-concordant fraction	0.671**	0.650**	0.655**	0.543**
Incorrect-discordant fraction	0.749**	0.726**	0.732**	0.647**
Correct-desensitized fraction	0.941	0.952	0.981	0.936*
Incorrect-desensitized fraction	0.708**	0.698**	0.681**	0.590**
Last-demonstration status				
Last demonstration not incorrect	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Last demonstration incorrect	0.991	1.041*	0.979	1.073**
Similarity and redundancy covariates				
Mean query-demo similarity, centered	1.886**	1.677**	0.990	2.356**
Prompt redundancy, per 1-unit increase	1.153*	0.847*	1.068	0.934
Query-group interactions				
Unprivileged-query × correct-discordant fraction	1.006	1.106*	1.146*	0.896*
Unprivileged-query × incorrect-concordant fraction	1.014	1.004	1.313**	1.310**
Unprivileged-query × incorrect-discordant fraction	0.977	0.964	1.179*	1.002
Unprivileged-query × correct-desensitized fraction	1.030	1.079	1.020	0.979
Unprivileged-query × incorrect-desensitized fraction	0.994	0.959	1.285**	1.142*
Unprivileged-query × last demonstration incorrect	0.978	0.992	1.015	0.972

Table 5: Full Harvard-FairVLMed fixed-effect odds ratios. Reference levels are shown in the table. Prompt-composition ORs replace the correct-concordant fraction. Stars mark * $p < .05$, ** $p < .001$. Dashes mark absent terms.

Term	Sex GPT-family OR	Sex Qwen-family OR	Race GPT-family OR	Race Qwen-family OR
Core covariates				
Intercept	4.786**	3.591**	3.417**	3.539**
Task difficulty, per 1-unit increase	0.218**	0.421**	0.304**	0.285**
Query group				
Privileged query	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Unprivileged query	0.945	1.097	0.805*	0.734**
Prompt count				
One demonstration	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Two demonstrations	0.982	1.037	1.051	0.989
Three demonstrations	0.997	0.982	1.063	0.973
Prompt-composition main effects				
Correct-concordant fraction	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Correct-discordant fraction	1.064	0.922	0.546**	0.486**
Incorrect-concordant fraction	0.195**	0.238**	0.184**	0.158**
Incorrect-discordant fraction	0.214**	0.269**	0.348**	0.380**
Correct-desensitized fraction	0.947	0.883	0.753**	0.728**
Incorrect-desensitized fraction	0.208**	0.278**	0.246**	0.232**
Last-demonstration status				
Last demonstration not incorrect	1 [Reference]	1 [Reference]	1 [Reference]	1 [Reference]
Last demonstration incorrect	1.185**	0.939*	1.113*	1.054
Similarity and redundancy covariates				
Mean query-demo similarity, centered	8.294	< 0.001**	2.326	123.890
Prompt redundancy, per 1-unit increase	–	–	–	–
Query-group interactions				
Unprivileged-query × correct-discordant fraction	0.991	0.842*	2.103**	2.070**
Unprivileged-query × incorrect-concordant fraction	0.988	0.963	2.508**	3.326**
Unprivileged-query × incorrect-discordant fraction	0.979	1.110	1.060	1.434*
Unprivileged-query × correct-desensitized fraction	0.997	0.979	1.481**	1.554**
Unprivileged-query × incorrect-desensitized fraction	1.112	1.030	1.549**	2.263**
Unprivileged-query × last demonstration incorrect	0.959	0.900*	0.991	0.866*

E Exploratory Noise-Stress Ablation Protocol

The stress-test ablations perturb displayed demonstration labels in the candidate pools, not query labels. Risk is the candidate-level estimate from Appendix H.1. We flip displayed labels until the row-weighted mean risk reaches the condition target: 0.40 for CheXpert and 0.50 for Harvard-FairVLMed.

Asymmetric group-noise injection applies flips only to the unprivileged demonstration group (Female for sex, Black for race); overall noise injection applies them pool-wide. For CheXpert, each demonstration has five labels, so at most two labels are flipped per demonstration, with flips balanced across abnormalities. For Harvard-FairVLMed, each flipped binary label changes positive to negative or negative to positive, and risk is recomputed from the flipped profile.

For Harvard-FairVLMed, original row-weighted risks were 0.358 and 0.386 for Female and Male rows and 0.468 and 0.360 for Black and White rows; affected groups reached 0.500 after injection. The same risk-achievement values apply to GPT-family and Qwen-family within each attribute.

The noise-stress settings also differ in visual similarity. Mean query-demonstration similarity was 0.353 for CheXpert race analyses and 0.394 for CheXpert sex analyses, versus 0.985 for Harvard-FairVLMed race analyses and 0.970 for Harvard-FairVLMed sex analyses. Thus the same stress protocol is evaluated in a lower-similarity chest-radiograph setting and a higher-similarity Harvard-FairVLMed setting.

F Label-Noise Injection Balanced Accuracy Figures

Figure 5 reports the macro balanced accuracy analogue to the main-text macro F1 score injection analysis.

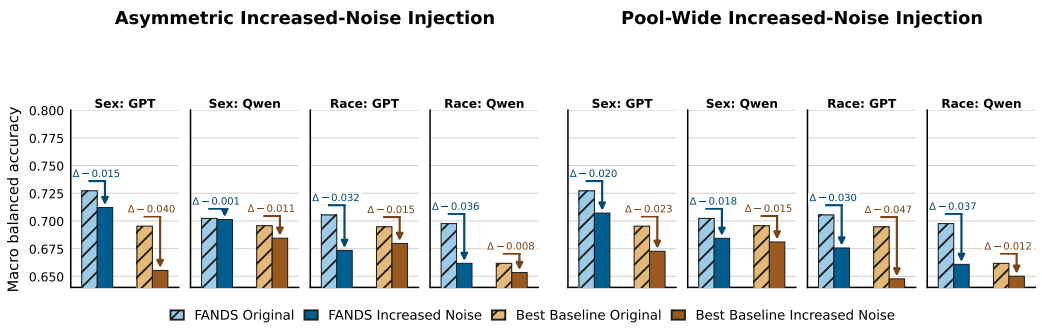

Figure 5: CheXpert 40% target-risk label-noise ablation on macro balanced accuracy. Injection settings match Figure 4(A). Arrows show post-injection minus original values.

The Harvard-FairVLMed balanced-accuracy analogue is shown in Figure 6.

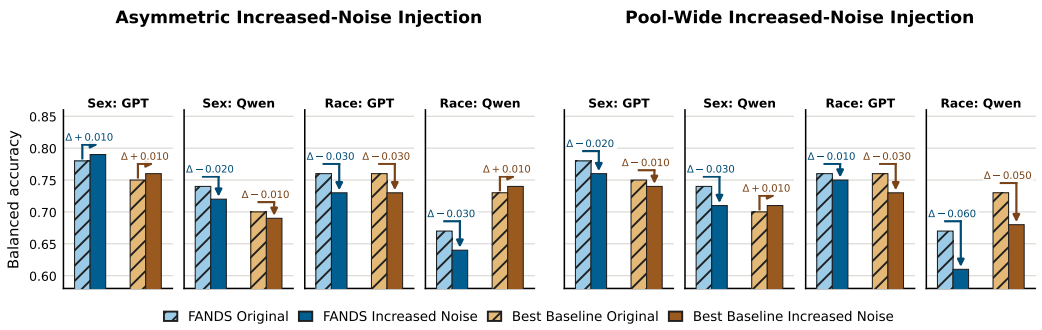

Figure 6: Harvard-FairVLMed 50% target-risk label-noise ablation on balanced accuracy. Injection settings match Figure 4(B). Arrows show post-injection minus original values.

G Data Processing

Existing data assets. CheXpert is distributed by Stanford AIMI for non-commercial research use, and Harvard-FairVLMed is distributed under CC BY-NC-ND 4.0 for non-commercial research. We use these existing datasets for experimental evaluation and do not introduce a new dataset asset.

Black-box LLMs. GPT-family uses proprietary `gpt-5` through the OpenAI API; its model size is not publicly disclosed. Qwen-family uses `qwen/qwen3-vl-235b-a22b-instruct` through OpenRouter, an instruction-tuned Qwen3-VL MoE model with about 235B total and 22B activated parameters.

Compute and API settings. Black-box LLM inference used API calls rather than local hosting or fine-tuning. Validation and test calls used temperature 0.0, a 32-token output cap, minimal reasoning effort, low response verbosity, a 90 s timeout, up to four attempts, single-worker execution, 0.5 s between calls, and capped retry backoff. We used prompt inputs and generated outputs only, without logits, hidden states, gradients, fine-tuning interfaces, or internal confidence scores. Local preprocessing, prompt construction, feature extraction, diagnostic regressions, surrogate training and selection, and metric computation ran on Ubuntu 22.04.5 with an AMD Ryzen 7 3700X CPU, 31

GiB RAM plus 15 GiB swap, and an NVIDIA RTX 3080 10 GiB GPU (driver 535.288.01, CUDA 12.2). For Harvard-FairVLMed, diagnostic regressions took about 15 minutes per analysis setting, or about 1 hour total; surrogate training, tuning, and testing took about 3–4 hours per analysis setting, or 12–16 hours across sex and race with GPT-family and Qwen-family analyses. CheXpert uses the same split sizes and model-selection protocol, so expected local compute is similar.

Table 6: Datasets and evaluation axes. The same prompt-composition vocabulary is used across datasets. Concordant, discordant, desensitized, correct, and incorrect demonstrations are defined relative to the current query and attribute.

Dataset	Prediction task	Labels used here	Attribute axes	Black-box LLMs
CheXpert	Chest-radiograph classification	Five binary findings	Sex, race	GPT-family, Qwen-family
Harvard-FairVLMed	Glaucoma classification	Binary target label	Sex, race	GPT-family, Qwen-family

The five CheXpert labels are Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. The selected-prompt splits are subject-disjoint. The CheXpert splits are balanced by the analyzed sensitive-attribute group, with 350/350 training subjects, 100/100 validation subjects, and 50/50 test subjects across the two groups. Within each group and split, one fifth of query subjects are no-finding cases and four fifths have at least one target finding.

The selected-prompt evaluations for both datasets use the 1,000-query split described in Section 6. It includes 700 query subjects for selection-surrogate training, 200 validation query subjects for method selection, and 100 held-out test query subjects for the final comparison. Prompt construction distinguishes whether the sensitive-attribute cue is shown or suppressed while preserving content and target-label evidence.

Candidate demonstration pool. For each validation or held-out query q, all methods use the same candidate demonstration pool $\mathcal{D}_q$, containing 50 eligible demonstrations drawn only from non-test data. The pool contains 25 demonstrations concordant with the query's sensitive-attribute group and 25 discordant demonstrations. Same-subject and, where applicable, same-study candidates are excluded. Neither held-out query labels nor held-out LLM responses are used in its construction.

For CheXpert, candidates within each group relation are collected using four deterministic rankings: query–demonstration similarity; similarity with an estimated-noise-risk penalty; MMR relevance–diversity ranking; and displayed-label coverage. Duplicates are merged, and a fixed greedy rule retains 25 candidates per relation by favoring similarity, lower estimated noise risk, coverage across the four rankings, displayed-label balance, and lower redundancy. All construction constants are fixed before held-out evaluation. For Harvard-FairVLMed, the 25 eligible demonstrations with the highest cosine similarity are retained within each relation.

Candidate prompt pool. The candidate prompt pool $\mathcal{P}_q$ contains every unique one-, two-, and three-demonstration combination formed from $\mathcal{D}_q$. The demonstrations retain their fixed order from $\mathcal{D}_q$; separate permutations are not enumerated. Each combination is represented in both cued and desensitized form. Thus, $|\mathcal{P}_q| = 2\left[\binom{50}{1} + \binom{50}{2} + \binom{50}{3}\right] = 41{,}750$ eligible prompts per query. FANDS scores every demonstration in $\mathcal{P}_q$. The baselines begin from the same $\mathcal{D}_q$ but apply their respective retrieval or search procedures.

H Method Details

H.1 Estimated Noise-Risk Feature

CheXpert embeddings use the public Microsoft RAD-DINO Hugging Face checkpoint (Pérez-García et al., 2025); Harvard-FairVLMed embeddings use fine-tuned FairCLIP checkpoints from the Harvard Ophthalmology AI Lab repository (Luo et al., 2024). These embeddings define query-demonstration similarity, nearest-neighbor estimated-noise-risk checks, and surrogate-model input features.

For candidate demonstration i, let $\mathcal{N}_k(i)$ be the set of $k = 15$ nearest examples in the original full candidate-demonstration pool after excluding disallowed same-subject or same-study neighbors. Let s_{ij} be cosine similarity between candidate i and neighbor j, and let $\tilde{y}_{i\ell}$ be the displayed label for label $\ell \in \{1, \ldots, C_d\}$. The neighbor-agreement score is

$$\hat{\eta}_i = \frac{\sum_{j \in \mathcal{N}_k(i)} \exp(\kappa s_{ij}) \left(C_d^{-1} \sum_{\ell=1}^{C_d} \mathbb{1}[\tilde{y}_{i\ell} = \tilde{y}_{j\ell}] \right)}{\sum_{j \in \mathcal{N}_k(i)} \exp(\kappa s_{ij})}, \qquad \kappa = 5. \tag{8}$$

The inner average is the fraction of labels on which candidate i and neighbor j agree. The exponential weight gives higher weight to more similar neighbors, and the denominator normalizes the weighted average. The unclipped risk is $1 - \hat{\eta}_i$; the clipped estimated noise-risk feature is

$$\hat{\epsilon}_i = \mathrm{clip}(1 - \hat{\eta}_i, 0.02, 0.98). \tag{9}$$

This quantity is computed at candidate-pool construction time and then used as the candidate-level risk input for prompt scoring. It is a selection-time proxy, not an audited probability that the displayed label is wrong.

H.2 Baseline Selection Rules

This appendix specifies the baseline selection rules used in Section 6. All baselines use the same image embeddings, candidate demonstration pool $\mathcal{D}_q$, and three-demonstration budget as FANDS. We implement EPR, EASE, and FCG for image–label demonstrations while retaining the central selection principle of each method. These three baselines are tuned separately for each dataset, sensitive attribute, and black-box LLM using only the training and validation splits. Held-out query labels and LLM responses are reserved for final evaluation, and these baselines do not use the FANDS estimated noise-risk feature.

Relevance-risk score The relevance-risk score ranks fixed three-demonstration prompts by combining average query-demonstration similarity with average estimated noise risk. For query q and candidate demonstration set $\pi = (i_1, i_2, i_3)$, let s_{qi} be the similarity between query q and candidate demonstration i, and let $\hat{\epsilon}_i$ be the estimated noise risk from H.1. The prompt-level averages are

$$\bar{s}_{q\pi} = \frac{1}{3} \sum_{i \in \pi} s_{qi}, \qquad \bar{\epsilon}_\pi = \frac{1}{3} \sum_{i \in \pi} \hat{\epsilon}_i. \tag{10}$$

Here $i \in \pi$ denotes the demonstration indices appearing in the ordered prompt π. The baseline score is

$$B(q, \pi) = \frac{1}{2} z_q(\bar{s}_{q\pi}) - \frac{1}{2} z_q(\bar{\epsilon}_\pi). \tag{11}$$

Here $z_q(\cdot)$ standardizes each prompt-level quantity over the available three-demonstration cued prompts for the same query. Larger values of $B(q, \pi)$ therefore correspond to prompts with higher average similarity and lower average estimated noise risk after within-query standardization.

Full-pool relevance-risk retrieval Full-pool relevance-risk retrieval selects the prompt with the largest $B(q, \pi)$ among all eligible three-demonstration cued prompts. This baseline represents a fixed high-similarity, low-risk retrieval rule that can draw from the full candidate pool.

Same-group relevance-risk retrieval Same-group relevance-risk retrieval applies the same score after restricting the prompt pool to demonstrations whose sensitive-attribute group matches the query group. This baseline represents a fixed same-group retrieval rule with the same relevance-risk tradeoff.

MMR diversity retrieval MMR diversity retrieval builds a three-demonstration prompt greedily. At step $t \in \{1, 2, 3\}$, let $\mathcal{A}_{t-1}$ be the demonstrations already selected. The next demonstration is chosen by

$$i_t = \arg \max_{i \in \mathcal{D}_q \setminus \mathcal{A}_{t-1}} \left[\lambda_{\mathrm{MMR}} s_{qi} - (1 - \lambda_{\mathrm{MMR}}) \max_{h \in \mathcal{A}_{t-1}} s_{ih} \right], \tag{12}$$

with the redundancy term set to zero when $\mathcal{A}_{t-1}$ is empty. Here s_{ih} is the similarity between two candidate demonstrations and $\lambda_{\mathrm{MMR}} \in [0, 1]$ controls the relevance-diversity tradeoff. This baseline does not use the estimated noise-risk score.

Visual EPR Following the learned-retrieval principle of EPR (Rubin et al., 2022), we use separate projection networks for the query image and each demonstration. The demonstration representation combines its image embedding with its displayed label, and the retrieval score is

$$r_\theta(q, i) = \tau_\theta \, f_q(z_q)^\top f_d([z_i; \widetilde{y}_i]), \tag{13}$$

where both representations are normalized and τ_θ is learned. We train the retriever on non-test one-demonstration evaluations with unmodified displayed labels, using the fraction of correctly predicted target labels to construct within-query ranking supervision. Architecture and training hyperparameters are selected on the validation split. At selection time, Visual EPR ranks the demonstrations in $\mathcal{D}_q$ and uses the three highest-scoring demonstrations, in score order, in a cued prompt.

Multimodal EASE Following the ordering-aware principle of EASE (Wu et al., 2024), we use a query encoder, position-specific demonstration encoders, and a reward network to predict the utility of an ordered three-demonstration sequence:

$$\widehat{u}_\theta(q, i_1, i_2, i_3) = \sigma(h_\theta([f_q(z_q); f_1(z_{i_1}, \widetilde{y}_{i_1}); f_2(z_{i_2}, \widetilde{y}_{i_2}); f_3(z_{i_3}, \widetilde{y}_{i_3})])) \, . \tag{14}$$

The reward network is trained on non-test three-demonstration evaluations with unmodified displayed labels. A query-bootstrap ensemble provides the mean and standard deviation of the predicted utility, and sequences are ranked by

$$S_{\mathrm{EASE}}(q, \pi) = \overline{u}(q, \pi) + \beta \, \mathrm{sd}(u(q, \pi)), \tag{15}$$

where β and the model hyperparameters are selected on the validation split. Multimodal EASE scores every ordered three-demonstration sequence from $\mathcal{D}_q$ and uses the highest-scoring sequence in a cued prompt.

FCG fairness-aware selection FCG (Hu et al., 2024) has two selection stages. First, we divide $\mathcal{D}_q$ by sensitive-attribute group and displayed-label status and apply K-means within each stratum, retaining demonstrations near the cluster centers. Second, a genetic search chooses three of these representatives. We estimate each demonstration's utility from non-test one-demonstration evaluations, with shrinkage toward the overall mean. Each three-demonstration set π is scored as

$$J_{\mathrm{FCG}}(q, \pi) = \alpha U(\pi) + (1 - \alpha)[w_g B_g(\pi; t_g) + w_y B_y(\pi) + w_d D(\pi) + w_s C(q, \pi)] \, , \tag{16}$$

where U is mean demonstration utility; B_g rewards proximity to the target unprivileged-group proportion t_g; B_y rewards displayed-label balance; D measures embedding diversity; and C is mean query–demonstration cosine similarity. We keep the group- and label-balance weights fixed across settings and select α, t_g, and the allocation between diversity and similarity on the validation split. The selected demonstrations form a cued prompt.

During held-out selection, the three learned and fairness-aware baselines also exclude demonstration-correctness indicators and corruption flags.

H.3 Selection Surrogate Details

The selection surrogate uses only information available before the held-out black-box LLM is queried. The model-ready feature families are as follows.

1. query context, including the sensitive-attribute group used for the current evaluation.
2. prompt length and prompt form, including whether the prompt uses cued or desensitized demonstrations.
3. group-relation patterns between the query and demonstrations, including same-group indicators.
4. query and selected-demonstration image embeddings, plus query-demonstration similarity features.
5. estimated noise-risk features for each selected demonstration and summary statistics across the candidate demonstration set.
6. displayed-label summary features and query-local label-prior summaries.

7. pairwise demonstration similarity features that summarize redundancy within the candidate demonstration set.

The feature design excludes held-out labels during selection, held-out LLM responses, oracle demonstration-correctness indicators as direct features, and per-slot desensitization indicators.

For query-local label priors, let $n^+_{q\ell}$ be the number of displayed positive labels among the six nearest eligible neighbors of query q in the full non-test demonstration pool. We use Jeffreys smoothing for this query-local scoring input.

$$\hat{\rho}_{q\ell} = \frac{n^+_{q\ell} + 0.5}{6 + 1}. \tag{17}$$

The added 0.5 pseudo-count prevents priors from being exactly zero or one with six neighbors. The resulting value is a query-local scoring input, not a population prevalence estimate. The exact surrogate architecture, optimization objective, training procedure, and setting-specific hyperparameters are reported in Appendix J.

H.4 Expected utility demonstration-set score

The main text defines FANDS as selecting the candidate demonstration set with the largest score $S(q, \pi)$. This score is an expected utility over candidate demonstration sets. This appendix gives the full construction.

The surrogate model estimates sensitivity and specificity conditional on the true label. For query q, candidate demonstration set π, and label ℓ, let $\hat{a}_{q\pi\ell}$ denote predicted sensitivity and $\hat{b}_{q\pi\ell}$ denote predicted specificity. The predicted sensitivity $\hat{a}_{q\pi\ell}$ estimates the probability that $\hat{y}_{q\ell}(\pi) = 1$ when $y_{q\ell} = 1$. The predicted specificity $\hat{b}_{q\pi\ell}$ estimates the probability that $\hat{y}_{q\ell}(\pi) = 0$ when $y_{q\ell} = 0$.

At selection time, the true label of the held-out query is unknown. We therefore estimate the probability that label ℓ is positive for query q. Let $\hat{\rho}_{q\ell}$ denote this query-specific label positivity estimate. Then $1 - \hat{\rho}_{q\ell}$ is the estimated probability that label ℓ is negative for query q.

We combine these probabilities with the predicted sensitivity and specificity to form expected classification terms.

$$\widehat{TP}_{q\pi\ell} = \hat{\rho}_{q\ell}\hat{a}_{q\pi\ell}, \qquad \widehat{FN}_{q\pi\ell} = \hat{\rho}_{q\ell}(1 - \hat{a}_{q\pi\ell}), \tag{18}$$

$$\widehat{TN}_{q\pi\ell} = (1 - \hat{\rho}_{q\ell})\hat{b}_{q\pi\ell}, \qquad \widehat{FP}_{q\pi\ell} = (1 - \hat{\rho}_{q\ell})(1 - \hat{b}_{q\pi\ell}). \tag{19}$$

Here $\widehat{TP}_{q\pi\ell}$ is the expected true-positive contribution, $\widehat{FN}_{q\pi\ell}$ is the expected false-negative contribution, $\widehat{TN}_{q\pi\ell}$ is the expected true-negative contribution, and $\widehat{FP}_{q\pi\ell}$ is the expected false-positive contribution. These are expected quantities computed before the black-box LLM response is observed.

The multiplication follows the probability product rule. For example, an expected true positive requires two events. Label ℓ is truly positive, and the black-box LLM predicts positive. The estimated probability of this joint event is the estimated probability that the label is positive, $\hat{\rho}_{q\ell}$, multiplied by the predicted probability that the LLM predicts positive given that the label is positive, $\hat{a}_{q\pi\ell}$.

Using the expected classification terms, we first define per-label expected accuracy, F1, sensitivity, and specificity.

$$\widehat{\mathrm{Acc}}_{q\pi\ell} = \widehat{TP}_{q\pi\ell} + \widehat{TN}_{q\pi\ell},$$

$$\widehat{\mathrm{F1}}_{q\pi\ell} = \frac{2\widehat{TP}_{q\pi\ell}}{2\widehat{TP}_{q\pi\ell} + \widehat{FP}_{q\pi\ell} + \widehat{FN}_{q\pi\ell}},$$

$$\widehat{\mathrm{Sens}}_{q\pi\ell} = \hat{a}_{q\pi\ell}, \qquad \widehat{\mathrm{Spec}}_{q\pi\ell} = \hat{b}_{q\pi\ell}.$$

Expected accuracy sums the expected true-positive and true-negative contributions, expected F1 uses the usual F1 denominator with expected classification terms, and expected sensitivity and specificity are the two surrogate-head predictions.

For multi-label tasks, the score macro-averages these quantities across the C_d labels in dataset d.

$$\widehat{\mathrm{Acc}}(q, \pi) = \frac{1}{C_d}\sum_{\ell=1}^{C_d} \widehat{\mathrm{Acc}}_{q\pi\ell},$$

$$\widehat{\mathrm{F1}}(q,\pi) = \frac{1}{C_d}\sum_{\ell=1}^{C_d} \widehat{\mathrm{F1}}_{q\pi\ell},$$

$$\widehat{\mathrm{Sens}}(q,\pi) = \frac{1}{C_d}\sum_{\ell=1}^{C_d} \widehat{\mathrm{Sens}}_{q\pi\ell}, \qquad \widehat{\mathrm{Spec}}(q,\pi) = \frac{1}{C_d}\sum_{\ell=1}^{C_d} \widehat{\mathrm{Spec}}_{q\pi\ell}.$$

The macro averages give equal label weight to each expected quantity.

For the optional hinge, we use the prior-weighted expected specificity guard.

$$\widehat{\mathrm{Spec}}_{\mathrm{guard}}(q,\pi) = \frac{\sum_{\ell=1}^{C_d}(1-\hat{\rho}_{q\ell})\hat{b}_{q\pi\ell}}{\sum_{\ell=1}^{C_d}(1-\hat{\rho}_{q\ell})}.$$

Jeffreys smoothing of $\hat{\rho}_{q\ell}$ keeps this denominator positive, and the numerical implementation also uses a small denominator floor for stability.

The selection score is the weighted expected utility with the optional specificity hinge.

$$\begin{aligned} S(q,\pi) = {} & w_{\mathrm{acc}}\widehat{\mathrm{Acc}}(q,\pi) + w_{\mathrm{f1}}\widehat{\mathrm{F1}}(q,\pi) \\ & + w_{\mathrm{spec}}\widehat{\mathrm{Spec}}(q,\pi) + w_{\mathrm{sens}}\widehat{\mathrm{Sens}}(q,\pi) \\ & - \lambda \max\{0, \tau - \widehat{\mathrm{Spec}}_{\mathrm{guard}}(q,\pi)\}. \end{aligned}$$

Here w_{acc}, w_{f1}, w_{spec}, and w_{sens} are nonnegative utility weights. The parameter $\lambda \geq 0$ controls the strength of the optional specificity hinge, τ is the target specificity level used by the hinge, and $\widehat{\mathrm{Spec}}_{\mathrm{guard}}(q,\pi)$ is the expected specificity quantity used for the optional guard. When $\lambda = 0$, the hinge is inactive.

H.5 Additional Diagnostic Findings

The main text keeps the diagnostic findings that directly motivate FANDS: concordance does not protect against label noise, desensitization is not a label-noise fix, and the relevant prompt-composition associations are setting-specific. The remaining coefficient families are supporting diagnostics for that third point.

Prompt count. Adding more demonstrations does not by itself produce a consistent fitted correctness gain. Across the eight diagnostic settings, two-demonstration prompts have odds ratios 0.976–1.051 and three-demonstration prompts have odds ratios 0.973–1.063. Neither prompt-count term is significant in any setting.

Query-group terms. The query-group coefficients support later subgroup evaluation rather than a coefficient-only fairness claim. Across the eight diagnostic settings, unprivileged-query main-effect odds ratios are 0.604–1.097, and they are significantly below the privileged-query reference in five settings, nonsignificant in three settings, and significantly above the reference in zero settings.

Control covariates and interactions. The final-position, similarity, redundancy, and query-group interaction rows are reported in Appendix D to make the diagnostic model auditable. They are not used as standalone evidence that a longer prompt, a desensitization rule, or a same-group rule is a fairness safeguard.

I Metric Definitions and Statistical Uncertainty

For CheXpert, we compute F1, balanced accuracy, sensitivity, and specificity separately for each of the five binary findings and macro-average them across findings. For Harvard-FairVLMed, we compute the corresponding binary metrics for its single binary endpoint.

For sensitive-attribute groups $\mathcal{G} = \{g_0, g_1\}$, we define worst-group F1 for selection policy m as

$$\text{WG-F1}_m = \min_{g\in\mathcal{G}} \mathrm{F1}_{m,g}\,. \tag{20}$$

Higher worst-group F1 indicates stronger performance for the lower-performing sensitive-attribute group.

Statistical uncertainty. We estimate uncertainty using bootstrap resampling at the subject level for held-out performance and at the candidate-demonstration level for estimated noise-risk audits. Subject-level resampling is stratified by sensitive-attribute group, and the same resampled subjects are used across selection policies in paired comparisons. Each CheXpert subject's five findings remain one resampling unit. When reporting uncertainty for an estimate or paired contrast, we use a symmetric 95% bootstrap confidence interval, with margin equal to the 95th percentile of the absolute bootstrap deviation from its observed point estimate.

J FANDS Selection Surrogate Architecture and Training

Model architecture. FANDS uses a supervised feed-forward multilayer perceptron (MLP). For query q and candidate demonstration set π, we denote its surrogate feature vector by $\mathbf{z}_{q,\pi}$. The feature families are defined in Appendix H.3. We fit all preprocessing operations on the training split: continuous features are median-imputed and standardized, and categorical features are one-hot encoded. A shared MLP backbone applies a linear transformation, optional batch normalization, the selected activation function, and dropout at each hidden layer.

Separate sensitivity and specificity heads produce $\widehat{a}_{q\pi\ell}$ and $\widehat{b}_{q\pi\ell}$, respectively, the definitions are in Section 5. CheXpert uses two five-output heads, with one output per finding. Harvard-FairVLMed uses two scalar outputs for binary glaucoma prediction.

For Harvard-FairVLMed, separate learned projections encode the 512-dimensional FairCLIP query and demonstration embeddings. For each active demonstration slot, the raw visual branch includes the projected query and demonstration embeddings, their element-wise product and absolute difference, the displayed demonstration label, age, group embedding, and slot-presence indicator. It also includes query demographics, demonstration-set size, and the cued/desensitized indicator. Desensitized demonstration groups are represented by a dedicated hidden category. We concatenate this representation with the structured surrogate features before passing it to the shared MLP.

Training objective. Let

$$c_{q\pi\ell} = \mathbf{1}\{\widehat{y}_{q\ell}(\pi) = y_{q\ell}\}$$

denote logged black-box LLM correctness for query q, candidate demonstration set π, and label ℓ. We train the sensitivity head only on observations for which $y_{q\ell} = 1$ and the specificity head only on observations for which $y_{q\ell} = 0$:

$$\mathcal{L}_{\text{cond}} = \frac{1}{2}\left[\mathbb{E}_{y_{q\ell}=1}\ell_\gamma(\widehat{a}_{q\pi\ell}, c_{q\pi\ell}) + \mathbb{E}_{y_{q\ell}=0}\ell_\gamma\left(\widehat{b}_{q\pi\ell}, c_{q\pi\ell}\right)\right].$$

For predicted probability p and binary target c, we define

$$p_t = cp + (1-c)(1-p), \qquad \ell_\gamma(p,c) = -(1-p_t)^\gamma \log p_t.$$

Thus, $\gamma = 0$ gives binary cross-entropy, while $\gamma \in \{1, 2\}$ gives the focal-loss configurations evaluated in our model search.

FANDS also learns to rank candidate demonstration sets within each query. Let $S(q, \pi)$ denote the expected-utility score defined in Appendix H.4 and let $u(q, \pi)$ denote the observed utility of a logged candidate demonstration set. For an ordered within-query pair satisfying $u(q, \pi_a) > u(q, \pi_b)$, define

$$\Delta u_{qab} = u(q, \pi_a) - u(q, \pi_b), \qquad \Delta S_{qab} = S(q, \pi_a) - S(q, \pi_b).$$

The pairwise ranking loss is

$$\mathcal{L}_{\text{rank}} = \frac{\sum \Delta u_{qab} \log\left(1 + \exp\left[-\frac{\Delta S_{qab}}{T_r}\right]\right)}{\sum \Delta u_{qab}},$$

where T_r is the ranking temperature. We sample at most 1,000 ordered pairs per query. CheXpert uses observed sample F1 with the selected specificity penalty as $u(q, \pi)$, whereas Harvard-FairVLMed uses observed binary correctness.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cond}} + \omega_r \mathcal{L}_{\text{rank}},$$

where ω_r controls the contribution of the ranking objective.

Optimization and model selection. We optimize the candidate models using AdamW or RAdam, clip the gradient norm at 1.0, train for at most 60 epochs, and apply early stopping with patience 9 and minimum improvement 10^{-5}. Historical validation determines the checkpoint and candidate shortlist. Validation LLM calls then select the final configuration lexicographically by mean label accuracy, F1, balanced accuracy, exact match, and specificity.

Table 7 reports the selected configurations. BN denotes batch normalization; WD/B_q gives weight decay and the number of queries per batch; ω_r/T_r gives the ranking-loss weight and temperature; and λ/τ gives the specificity-penalty weight and threshold. Utility weights are ordered as $(w_{\mathrm{acc}}, w_{\mathrm{f1}}, w_{\mathrm{spec}}, w_{\mathrm{sens}})$.

Table 7: Validation-selected FANDS surrogate configurations.

Setting	Architecture	Trainable parameters	Loss	Optimizer/LR
CheXpert				
Sex / GPT-family	MLP 512–256–128/GELU/BN	195,978	Focal ($\gamma = 1$)	RAdam/.001
Sex / Qwen-family	MLP 256–128–64/GELU	56,138	Focal ($\gamma = 2$)	AdamW/.002
Race / GPT-family	MLP 256–128/GELU	48,522	BCE	RAdam/.002
Race / Qwen-family	MLP 256–128–64/GELU	56,138	Focal ($\gamma = 2$)	AdamW/.002
Harvard-FairVLMed				
Sex / GPT-family	FairCLIP 512→128; MLP 128–64/GELU	317,162	Focal ($\gamma = 2$)	RAdam/.002
Sex / Qwen-family	FairCLIP 512→96; MLP 256–128/SiLU	403,946	Focal ($\gamma = 1$)	AdamW/.003
Race / GPT-family	FairCLIP 512→96; MLP 256–128/GELU	403,946	Focal ($\gamma = 2$)	RAdam/.002
Race / Qwen-family	FairCLIP 512→96; MLP 256–128/SiLU/BN	404,714	BCE	AdamW/.001

Setting	WD/B_q	Dropout	ω_r/T_r	λ/τ	Utility weights
CheXpert					
Sex / GPT-family	.0001/8	.20	1.0/.20	.10/.80	(0,1,0,0)
Sex / Qwen-family	.001/16	.10	1.0/.10	.00/.82	(0,1,0,0)
Race / GPT-family	.001/8	.10	.5/.10	.10/.80	(0,1,0,0)
Race / Qwen-family	.001/16	.10	1.0/.10	.00/.82	(0,1,0,0)
Harvard-FairVLMed					
Sex / GPT-family	.003/8	.20	1.0/.20	.30/.82	(.35,.20,.10,.35)
Sex / Qwen-family	.003/8	.30	.2/.20	.20/.80	(.25,.25,.25,.25)
Race / GPT-family	.003/8	.20	1.0/.20	.20/.80	(.35,.20,.35,.10)
Race / Qwen-family	.003/16	.30	1.0/.05	.10/.75	(.35,.20,.35,.10)

K Estimated Noise-Risk Audits

We examine two properties of the estimated noise-risk feature $\widehat{\epsilon}_i$. First, we test whether it systematically assigns higher risk or provides weaker corruption discrimination for unprivileged groups. Second, we test whether FANDS remains effective when its candidate-level risk estimates are perturbed. Both analyses use the estimator defined in Appendix H.1 and leave the reported FANDS method unchanged.

K.1 Subgroup Behavior of Estimated Noise Risk

A systematic upward shift in $\widehat{\epsilon}_i$ for unprivileged demonstrations could discourage their selection. We therefore compare mean estimated risk across sensitive-attribute groups. Because neither dataset provides independently adjudicated mislabel annotations, we also conduct a paired controlled-corruption audit. For each candidate demonstration, we recompute $\widehat{\epsilon}_i$ after inverting its displayed labels while holding its image, neighborhood, and similarity weights fixed.

The paired detection rate measures how often the corrupted version receives higher estimated risk, with ties assigned one-half credit. AUROC measures discrimination between unmodified and inverted-label versions using $\widehat{\epsilon}_i$, treating inversion as the positive condition.

Table 8: Subgroup response of $\widehat{\epsilon}_i$ to controlled label corruption. AUROC is reported as point estimate $\pm$ symmetric 95% bootstrap margin; U and P denote unprivileged and privileged groups.

Dataset	Attribute	Group	Role	n	Mean unmodified $\widehat{\epsilon}_i$	Mean corrupted $\widehat{\epsilon}_i$	Paired detection	AUROC
CheXpert	Sex	Female	U	28,225	.28	.72	.87	$.94 \pm .00$
		Male	P	35,308	.27	.73	.88	$.95 \pm .00$
	Race	Black	U	2,591	.26	.74	.89	$.95 \pm .01$
		White	P	35,607	.29	.71	.87	$.94 \pm .00$
Harvard-FairVLMed	Sex	Female	U	3,115	.36	.64	.72	$.79 \pm .02$
		Male	P	2,146	.37	.63	.71	$.78 \pm .02$
	Race	Black	U	340	.45	.55	.56	$.61 \pm .06$
		White	P	4,465	.37	.63	.72	$.79 \pm .01$

Table 8 does not show a systematic increase in estimated risk for unprivileged groups. Sex differences are small in both datasets, and Black demonstrations receive lower mean risk than White demonstrations in CheXpert. For race in Harvard-FairVLMed, the higher mean estimated risk and weaker corruption discrimination for Black demonstrations coincide with sparse same-group neighborhood support.

Only 12% of the neighbors of Black demonstrations in Harvard-FairVLMed belong to the same group, compared with 93% for White demonstrations. Balancing group representation in the neighborhood raises Black-group AUROC from 0.61 to 0.73, and restricting the diagnostic neighborhood to the same group raises it to 0.84. These diagnostic results identify sparse local support as the source of this isolated disparity; the reported FANDS method remains unchanged.

The estimator is also stable across $k \in \{5, 10, 15, 20, 30\}$ and $\kappa \in \{0, 2.5, 5, 10\}$. Moreover, FANDS does not use $\widehat{\epsilon}_i$ as an exclusion threshold. Unprivileged demonstrations account for 40.3–61.2% of selected demonstration slots across the eight settings, indicating that estimated noise risk does not cause FANDS to systematically filter them out.

K.2 Sensitivity to Estimated Noise-Risk Error

We next test whether FANDS depends on precise candidate-level values of $\widehat{\epsilon}_i$. For each unique candidate demonstration i, we draw one fixed $xi_i \sim \mathrm{Uniform}(-1, 1)$ and define

$$\widehat{\epsilon}_i(\delta) = \mathrm{clip}\left(\widehat{\epsilon}_i + \delta u_i,\ 0,\ 1\right), \qquad \delta \in \{0.10, 0.15, 0.20\}.$$

The same xi_i is used at every perturbation level, making the perturbations nested.

We hold the validation-selected checkpoints, hyperparameters, query priors, embeddings, demonstration pool, held-out queries, and black-box LLM configurations fixed. We then recompute all features derived from $\widehat{\epsilon}_i$, rescore every candidate demonstration set in $\mathcal{P}_q$, and repeat selection without retraining or retuning FANDS. We evaluate the new selection with the corresponding black-box LLM.

Table 9: Sensitivity of FANDS to additive error in $\widehat{\epsilon}_i$, averaged across eight settings. Changes are percentage points relative to unperturbed FANDS and are reported as point estimate $\pm$ symmetric 95% bootstrap margin.

Risk perturbation	F1	Change in F1	Balanced accuracy	Change in balanced accuracy	F1 wins vs. fixed baseline
0%	.69	–	.72	–	7/8
10%	.68	-1.15 ± 1.22	.72	-0.74 ± 1.04	7/8
15%	.69	-0.22 ± 1.35	.72	$+0.04 \pm 1.08$	7/8
20%	.69	$+0.14 \pm 1.47$	.73	$+0.37 \pm 1.22$	7/8

F1 remains between 0.68 and 0.69, balanced accuracy remains between 0.72 and 0.73, and every paired performance interval includes zero. FANDS also continues to outperform the strongest fixed baseline in F1 in seven of eight settings at every perturbation level. These results show that FANDS benefits from estimated noise risk without requiring exact candidate-level values or a perfect risk ranking.

L Component-Level Ablations

We evaluate five component-ablation variants to determine which parts of FANDS contribute to its held-out performance. The first four variants remove one feature block, while the fifth replaces the output structure of the selection surrogate.

Ablated components. The *estimated noise-risk* block contains the estimated noise risk $\widehat{\epsilon}_i$ of each demonstration, demonstration-set risk summaries, and group-conditioned risk summaries. The *order-sensitive* block contains slot-presence indicators and position-indexed similarity, estimated noise risk, group relation, and inter-demonstration similarity. For Harvard-FairVLMed, removing this block also replaces the ordered raw demonstration representations with permutation-invariant pooling.

The *similarity/redundancy* block contains query–demonstration similarity, demonstration-set similarity summaries, and pairwise demonstration similarity. For Harvard-FairVLMed, this ablation also removes the raw embedding interactions from which these similarity signals could otherwise be recovered. The *group-relation* block contains slot-level concordant/discordant indicators, the corresponding counts and fractions, and group-conditioned risk summaries.

The architectural ablation retains all input features and the shared backbone but replaces the label-conditional sensitivity and specificity heads with a single scalar utility output. The order-sensitive block does not include the diagnostic last-demonstration-incorrect indicator, which is not used by FANDS during selection.

Evaluation protocol. We retrain every variant using the same subject-level splits and query-specific candidate demonstration pools as complete FANDS. We hold the non-ablated inputs, shared backbone, optimizer, ranking objective, early-stopping procedure, and random seed fixed. Each model scores the same eligible one-, two-, and three-demonstration configurations, and each selected configuration is evaluated using an actual held-out call to the corresponding black-box LLM. The displayed analysis contains five ablations for each of the four GPT-family dataset–attribute settings, with complete FANDS as the reference.

Figure 3 reports F1 in the main text. Figure 7 reports the corresponding balanced-accuracy results.

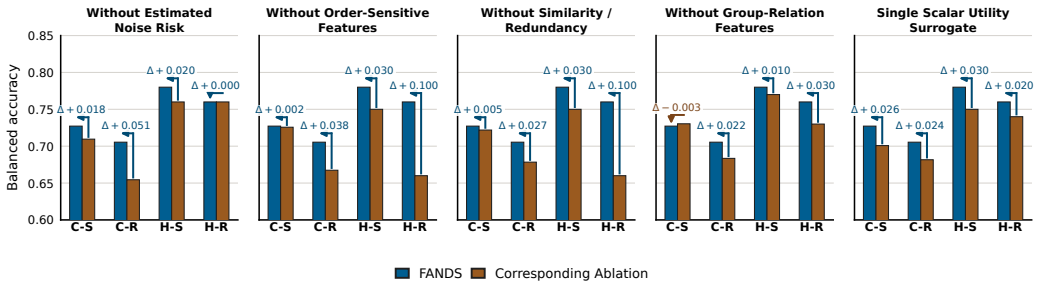

Figure 7: GPT-family component ablation on held-out balanced accuracy. Each panel compares FANDS with the ablation named in its heading. Arrows report $\Delta = \mathrm{FANDS} - \mathrm{ablation}$. C-S denotes CheXpert sex, C-R CheXpert race, H-S Harvard-FairVLMed sex, and H-R Harvard-FairVLMed race.

Across F1 and balanced accuracy, complete FANDS is best or tied in seven of the eight GPT-family metric–setting comparisons. The only exception is CheXpert sex balanced accuracy, where removing group-relation features increases performance by 0.003. Averaged across the four settings, retaining every component has a positive contribution: mean F1 improvements range from 2.40 to 4.83 percentage points, and mean balanced-accuracy improvements range from 1.48 to 4.23 percentage points. The results therefore show that no individual feature block or single scalar utility surrogate accounts for the performance of complete FANDS across datasets and sensitive attributes.

M SAMPLE EFFICIENCY

The primary experiments use 70,000 logged query–prompt outcomes per setting, obtained from 700 training subjects and 100 prompts per subject. To assess sample efficiency, we retrain each FANDS selection surrogate from scratch using a deterministic subset of 300 training subjects. The subset is balanced by the relevant sensitive attribute and selected to approximate the label prevalences of the complete training cohort. This produces 30,000 logged outcomes per setting, reducing the training budget by 57.1%.

We retain the validation-selected training configuration, validation and held-out splits, and query-specific candidate demonstration pools of the corresponding 70,000-outcome model. We then perform demonstration selection and obtain one black-box LLM response for each of the same 100 held-out subjects.

Table 10: Held-out FANDS performance using 70,000 and 30,000 logged query–prompt outcomes per setting. Values are point estimates ± symmetric 95% subject-bootstrap margins. Changes are 30k minus 70k in percentage points and use paired symmetric 95% bootstrap margins.

Dataset	Attribute	LLM	F1			Balanced accuracy		
			70k	30k	Change (pp)	70k	30k	Change (pp)
CheXpert	Sex	GPT-family	$.68 \pm .06$	$.66 \pm .06$	-1.75 ± 3.71	$.73 \pm .04$	$.72 \pm .05$	-0.93 ± 2.66
		Qwen-family	$.66 \pm .07$	$.64 \pm .07$	-2.21 ± 4.17	$.70 \pm .05$	$.70 \pm .05$	-0.57 ± 3.18
	Race	GPT-family	$.63 \pm .07$	$.61 \pm .07$	-2.69 ± 3.80	$.71 \pm .05$	$.68 \pm .05$	-2.30 ± 2.73
		Qwen-family	$.67 \pm .07$	$.63 \pm .07$	-4.26 ± 4.17	$.70 \pm .05$	$.67 \pm .05$	-2.72 ± 3.28
Harvard-FairVLMed	Sex	GPT-family	$.78 \pm .09$	$.73 \pm .10$	-4.53 ± 7.14	$.78 \pm .08$	$.74 \pm .09$	-4.00 ± 6.08
		Qwen-family	$.70 \pm .11$	$.70 \pm .11$	$+0.56 \pm 9.90$	$.74 \pm .08$	$.73 \pm .08$	-1.00 ± 7.62
	Race	GPT-family	$.76 \pm .09$	$.68 \pm .11$	-8.51 ± 7.11	$.76 \pm .08$	$.67 \pm .09$	-9.00 ± 7.11
		Qwen-family	$.63 \pm .12$	$.61 \pm .12$	-1.74 ± 8.55	$.67 \pm .09$	$.67 \pm .09$	0.00 ± 6.96
Equal-setting mean			$.69 \pm .03$	$.66 \pm .03$	-3.14 ± 2.29	$.72 \pm .02$	$.70 \pm .03$	-2.57 ± 1.91

The 30,000-outcome models retain 95.4% of the complete models' mean F1 and 96.5% of their mean balanced accuracy. Although the complete training budget provides a measurable average benefit, the reduced-budget models retain positive F1 margins over the baselines evaluated in this analysis. Across all eight settings, the one-time training budget decreases from 560,000 to 240,000 outcomes. After training, FANDS scores the eligible demonstration configurations locally and requires only one black-box LLM call per new query.

N CANDIDATE-SIMILARITY SENSITIVITY

Candidate demonstration pools differ substantially in query–demonstration similarity across datasets. To test whether FANDS depends on the original CheXpert relevance regime without comparing cosine scales across encoders, we conduct a controlled sweep. For each held-out query, we rank eligible non-test demonstrations separately within each sensitive-attribute group and construct four 50-demonstration pools from contiguous blocks beginning 10%, 25%, 50%, or 75% down each ranking. Each pool contains 25 demonstrations from each group, and the same-subject and same-study exclusions are retained.

We keep the validation-selected FANDS checkpoints, expected-utility parameters, and baseline rules fixed. FANDS and the baselines then select demonstrations for the same 100 held-out subjects in each of the four CheXpert attribute–LLM settings, followed by evaluation with the corresponding black-box LLM.

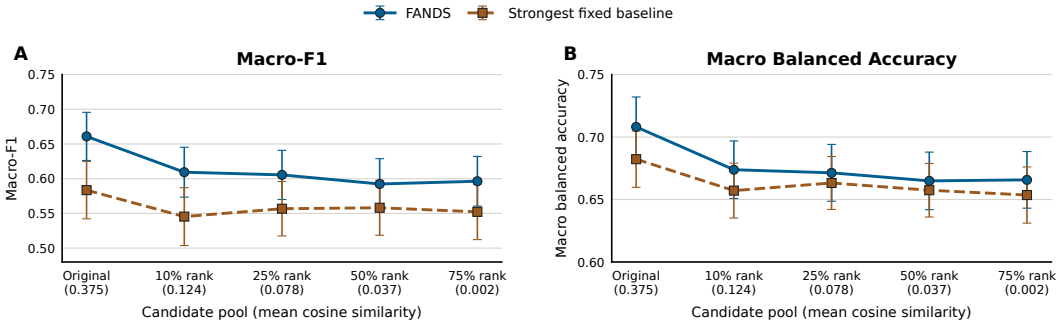

Figure 8: Candidate-similarity sensitivity on CheXpert. Points are equal-setting means across the four attribute–LLM settings, and error bars are symmetric 95% subject-bootstrap intervals. At each candidate set and metric, the comparison curve reports the strongest of the baselines. Parentheses on the horizontal axis give mean query–demonstration cosine similarity.

Absolute performance generally decreases as the available demonstrations become less relevant. Nevertheless, FANDS retains the highest equal-setting mean Macro-F1 and macro balanced accuracy at every tested similarity level. Relative to the strongest fixed baseline, its Macro-F1 advantage ranges from 3.43 to 7.74 percentage points, while its balanced-accuracy advantage ranges from 0.75 to 2.58 points. Together, these results show that FANDS remains robust across widely different candidate-similarity regimes.