

Colloquial Adverse Drug Reactions Are Consistently Harder: A Register-Stratified Diagnosis of Social-Media Extraction

Tomi Thomo¹ and Alex Thomo^{2*}

¹European University of Tirana, Tirana, Albania.

^{2*}Department of Computer Science, University of Victoria, Victoria, BC, Canada.

*Corresponding author(s). E-mail(s): thomo@uvic.ca;

Abstract

Patients describe medication side effects in online forums and reviews, a valuable but noisy pharmacovigilance source. Large language models can extract these adverse drug reactions from a few in-prompt examples, without task-specific training, and benchmarks report how well they do in a single score. That score hides a large disparity inside every model we tested. Sorting reactions by how they are worded, we find that those named in clinical terms are recovered about twice as often as those phrased in everyday, figurative language (span recall about 72% against 34% on psychiatric drug reviews). The disparity is not an artifact of prompting. A supervised biomedical encoder, fine-tuned on the gold spans, lifts both registers but leaves a comparable gap, and a paired comparison on the same items detects no recall advantage for it in either register; its aggregate lead rests principally on suppressing false positives. Nor is it confined to social media. It holds for four models and, separately, for four corpora, including clinician-written case reports, where colloquial phrasing is rare but no easier to extract, and it is not explained by span length or word rarity. As an instrument we build a training-free, CPU-only policy that guarantees scarce reaction-bearing examples a place in the prompt. It lifts colloquial recall to 51% while holding precision on reaction-bearing posts, at the cost of more false alarms on reaction-free ones. Stratifying a benchmark by how its target is expressed turns “fine-tuning wins” into a map of where each method helps.

Keywords: adverse drug reaction, pharmacovigilance, in-context learning, large language models, demonstration selection, social media mining

1 Introduction

People routinely describe the side effects of their medications in online forums and product reviews, often long before those effects surface in formal reporting channels. Collected at scale, these first-hand reports are a promising complement to traditional pharmacovigilance, the monitoring of drug safety after a product reaches the market (Sarker et al. 2015; Nikfarjam et al. 2015). The text is messy, though, with reactions named in passing, mixed with unrelated complaints, and often described in ordinary rather than clinical language. Turning such posts into structured safety signals requires reliably pulling out the span of text that names each adverse drug reaction (ADR).

Two families of methods dominate this extraction task. One fine-tunes a supervised model on human-labelled spans; the other prompts a large language model with a few worked examples placed directly in the input, so-called in-context learning, which needs no gradient updates and no task-specific training data (Brown et al. 2020; Dong et al. 2024). The second option is appealing precisely because labelled ADR spans are expensive to produce. On the standard benchmarks, however, the few-shot prompting route consistently lands well below the fine-tuned model in F1 (Elbiach et al. 2025; Gu et al. 2023). That gap is usually reported as a single aggregate number, and the natural reading is that the prompted model is simply a weaker extractor across the board.

We argue that this reading is wrong, or at least incomplete, and that the aggregate number hides structure that matters. Adverse reactions are described in two broadly different registers. Some are stated with clinically recognizable names (“nausea”, “insomnia”, “dizziness”); we call these *explicit*. Others are conveyed through everyday or figurative descriptions (“zoned out”, “not as sharp mentally”, “disconnected from my body”) that carry the same clinical meaning without naming a symptom; we call these *implicit*.

Our central question is therefore not *how large* the in-context-learning-versus-fine-tuning gap is, but how a single model’s accuracy varies between these two registers. That within-model question comes first. Without that stratification, a paradigm comparison collapses two populations that behave differently into a single number, and cannot show which of them drives the result. Our claim is that every model we test, prompted or fine-tuned, is markedly weaker on the implicit register, a disparity that aggregate F1 hides.

Showing that the disparity exists needs only stratified evaluation. Asking whether anything closes it without training needs an intervention to apply, and one is close at hand. A recurring detail in the benchmark literature is that the in-prompt examples are chosen *at random* from the training pool (Elbiach et al. 2025), which is a weak baseline for a method whose entire signal comes from those examples. We therefore build a training-free example-selection policy that replaces random choice with two ingredients, similarity-based retrieval so the examples resemble the post being labelled, and a minority guarantee so that reaction-bearing examples reach the prompt even when they are rare. Neither ingredient needs test-time labels, extra annotation, or gradient training. Retrieval is purely lexical, and the guarantee reads only the answers already attached to the candidate demonstrations, which few-shot prompting supplies anyway. Both run on a CPU. This selector is primarily a probe rather than an end

in itself. It lets us ask whether better example selection closes the register gap, and if not, where it stalls.

The answer, developed over the rest of the paper, is fourfold. First, on posts that do contain a reaction, the minority guarantee improves recall over strong retrieval while holding precision (and improving both on explicit spans), and the improvement concentrates on the implicit register where it is most needed. The effect is small but consistent, pointing the same way and by a similar amount on all four models and again under a register labeller from a different model family; it is essentially invisible in aggregate F1 because it is offset by a measurable rise in false alarms on reaction-free posts, which we report as an operating-point trade. Second, a large residual gap between explicit and implicit recall survives every training-free selection method we try.

Third, and most importantly, that residual is not an artifact of prompting. A supervised encoder fine-tuned directly on the gold labels still leaves a comparable explicit-versus-implicit gap and has a similar implicit-recall point estimate to the training-free selector, with no detectable difference between them in a paired comparison. Fine-tuning lifts the whole curve and sharply reduces false alarms, but it does not close the implicit deficit. The deficit therefore persists across every model and paradigm we evaluated, rather than being a property of any single one.

Fourth, it is not peculiar to the two benchmarks we start from. Applying the same stratification to four corpora, including biomedical case reports written by clinicians rather than patients, we find that how *often* reactions are phrased colloquially varies more than threefold, while how much harder they are to extract once phrased that way does not measurably differ across the four.

Our contributions are:

- A *stratified diagnosis* of social-media ADR extraction, showing that within any single model accuracy is far lower on implicitly expressed reactions than on clinically named ones, a disparity aggregate F1 hides entirely (Section 4.3). We keep this claim distinct from the separate questions of how prompting compares with fine-tuning and of how much demonstration selection adds.
- A *training-free, CPU-only demonstration-selection policy* (retrieval plus a minority guarantee) used as the instrument for the diagnosis, with an operating-point analysis of its recall benefit on reaction-bearing posts against its false-alarm cost on reaction-free posts (Sections 3 and 4.4).
- A *fine-tuned upper-bound reference line* (two supervised biomedical encoders, five seeds each) demonstrating that the implicit residual persists across the fine-tuning boundary, so the gap is not merely a prompting artifact (Section 4.5).
- A *cross-corpus test* of the diagnosis on four corpora spanning patient forums, drug reviews and biomedical case reports, which finds no detectable difference in the register penalty across them even though their absolute difficulty and their share of colloquial phrasing both vary widely (Section 6).
- A set of *validity checks* establishing that the diagnosis is not an artifact of how the labels were drawn or of what the spans look like: a human-validated register labeller with reported agreement, a train/test leakage audit, a boundary-sensitivity check, and confound-matching against span length and word rarity (Section 5).

2 Related Work

2.1 In-Context Learning and Demonstration Selection

In-context learning lets a language model perform a task from a few input-output examples supplied in the prompt, without parameter updates (Brown et al. 2020; Dong et al. 2024). Its behaviour is famously sensitive to which examples are shown and in what order (Lu et al. 2022; Min et al. 2022), which has motivated a line of work on choosing examples deliberately rather than at random. Retrieval-based selection picks examples similar to the test input, typically with a dense sentence encoder (Liu et al. 2022; Rubin et al. 2022). On biomedical event extraction specifically, a lexical scorer (BM25) has been found to beat both random selection and dense retrieval (Sun et al. 2024), consistent with the intuition that entity-bearing tasks reward surface overlap.

A second thread targets label structure and long-tail difficulty. Label-aware retrieval learns representations that separate adjacent, easily confused labels and helps most in the extreme few-shot regime (Chen et al. 2024). Uncertainty-based ranking up-weights hard or long-tail examples and reports gains on long-tail question answering, while cautioning that retrieval can also mislead the model on exactly those hard cases (Yu et al. 2025). Structure-based selection retrieves examples with complementary syntactic and semantic form and transfers across model families (Chakma et al. 2026), and fine-tuning a model specifically to exploit its few-shot examples has also been explored (Walsh et al. 2025).

Most directly, Peng et al. (2024) revisit demonstration selection itself and show that the best choice of examples is both data-dependent and dependent on the inference model, proposing a model-aware reranking that unifies several earlier selection heuristics.

That finding bears on our design in two ways. It suggests that no selector should be assumed to transfer unchanged across models and corpora, which is why we evaluate the same selector across four models spanning three families and report where it is flat as well as where it helps; and it supports reading a selector as a probe of what a given model is missing rather than as a universal improvement, which is the use we make of ours. Our selector borrows the label-aware idea from this thread but in its cheapest possible form. The minority class is guaranteed a place in the prompt by reading the gold answers already attached to the candidate demonstrations, so it needs no trained retriever, no extra annotation, and no test-time labels; and, crucially, we use it as a diagnostic instrument rather than proposing it as a new state of the art.

2.2 Adverse Drug Reaction Extraction from Social Media

Mining ADRs from user-generated text has a long history, from sequence labelling with engineered features (Nikfarjam et al. 2015) through the SMM4H shared-task series (Magge et al. 2021), where extraction is typically followed by normalization to a standard terminology (Yazdani et al. 2023). Our primary corpora are CADEC, forum posts with span-level ADR annotations (Karimi et al. 2015), and PsyTAR, patient reviews of psychiatric medications (Zolnoori et al. 2019). We add CADECv2 (Dai et al. 2024), an independently constructed patient-forum corpus, and PHEE (Sun et al.

2022), a biomedical-literature event-extraction dataset. PHEE is clinician-written rather than patient-authored, and we include it precisely for that contrast, to test whether the pattern we report is confined to social media (Section 6).

Whether adverse-event extraction survives a change of domain is itself an active question. CADECv2 reaches us through MultiADE (Dai et al. 2024), a benchmark assembled to study exactly that, across corpora that differ in source and annotation. That work asks how a *trained extractor* transfers; we ask a narrower question about the data, whether a particular linguistic difficulty travels with it, which is why Section 6 holds the model and the selection strategy fixed and varies only the corpus.

Two findings from recent work frame ours. The benchmark we reproduce and extend (Elbiach et al. 2025) reports the aggregate gap we decompose, and draws its few-shot examples at random. Its own qualitative error analysis already names implicit and colloquial reactions, negation, and multi-token spans as failure modes, but never quantifies them. Separately, generative extraction of ADR spans is known to be fragile. Shared-task systems have abandoned LLM extraction after observing hallucinated spans (Li et al. 2024), and a direct comparison found GPT-4 span-extraction F1 far below a supervised classifier on the same data (Zheng et al. 2024). Document-level and distillation studies reach the same ordering, with fine-tuned or distilled encoders on top (Gu et al. 2023; D’Oosterlinck et al. 2023; Harris et al. 2024). A psychiatric-ADR benchmark on social media is the nearest in spirit to our corpus choice and reports high false-positive rates on non-standard reactions (Chandra et al. 2025).

A structural feature of this data matters for our method. Reaction-bearing posts are often the minority. A systematic review of 51 studies across more than 174 social-media sites, most of them discussion forums, puts the prevalence of adverse-event reports at 0.2 to 8% of posts (Golder et al. 2015); consistent with this, fewer than one post in ten reports a reaction on the SMM4H 2019 Twitter set (Breden and Moore 2020), and even on review corpora the positive rate is well under half. Random exemplar selection therefore frequently fills a prompt with reaction-free posts, giving the model no worked example of the very thing it must extract. This is precisely the hole a minority guarantee is meant to fill, and it is why we expect (and find) the guarantee to help on the positive-scarce corpus and not on the positive-abundant one.

That colloquial reactions form a register distinct from clinical terminology is well documented in the consumer-health-vocabulary literature, which studies the systematic gap between lay expressions and their clinical equivalents (Zeng and Tse 2006); we quantify how that gap manifests as an extraction deficit. Across this literature, two strands remain loose, a named-but-unquantified implicit failure mode and an encoder advantage over prompted generative models. We tie them together by asking whether the encoder’s advantage is uniform or, like the prompted model’s deficit, concentrated on the implicit register.

2.3 Fine-Tuned Baselines and the Comparability Question

A natural objection to any “in-context learning trails fine-tuning” study is: which fine-tuned model? We use supervised biomedical encoder taggers (primarily PubMedBERT (Gu et al. 2021)) as the extraction upper bound, for a concrete reason. Generative LLM fine-tuning for this task is known to underperform and to hallucinate

spans (Li et al. 2024; Zheng et al. 2024), whereas encoder token-classification is the standard high-water mark for span extraction.

That fine-tuning leads in-context learning on this task is established by the peer-reviewed anchor and distillation studies (Elbiach et al. 2025; Gu et al. 2023; D’Oosterlinck et al. 2023), and a recent preprint reports the same ordering on CADEC across encoder and GPT-4o settings (Baroian 2025). A recent published ADR paper (Li and Lin 2026) fine-tunes a generative model with chain-of-thought prompting (Wei et al. 2022) and low-rank adaptation (Hu et al. 2022; Dettmers et al. 2023), but targets *binary detection* (does a post contain an ADR?), a different and easier task than extracting the reaction span. Its high detection F1 is therefore not comparable to span-extraction F1, and we never present our numbers as competing with it; we use it only as the source of one of our training-free selection baselines (its clustering-based selector) that we ablate.

The comparison is also task-dependent rather than uniform. In a broad GLUE-based study, Zhong et al. (2023) (a preprint) find that ChatGPT trails fine-tuned BERT-style models on paraphrase and similarity, outperforms them on inference, and is comparable on sentiment analysis and question answering, with improved prompting raising its performance further. This is the general form of the pattern we report. The interesting question is not whether prompting trails fine-tuning on average, but on which linguistic phenomena it does so.

2.4 The Gap We Fill

No prior work, to our knowledge, stratifies the in-context-learning-versus-fine-tuning gap on social-media ADR extraction by how the reaction is expressed, tests a training-free minority-aware selector on that stratification, checks whether a supervised upper bound closes the residual, and then asks whether the resulting deficit is particular to the corpora it was measured on. That decomposition, rather than the selector itself, is our contribution.

3 Methods

3.1 Task and Metrics

The task is ADR span extraction. Given a post, the goal is to return the set of text spans that name adverse reactions. Following the benchmark we reproduce (Elbiach et al. 2025), the model emits a comma-separated list of reaction spans, which we align to the gold spans and score with strict and relaxed F1. Strict F1 counts a prediction correct only on an exact span match; relaxed F1 counts a partial match (embedding cosine similarity above 0.8, following the anchor) as correct. Strict F1 is the clean, embedder-independent comparison and is our primary metric throughout; we report relaxed F1 only where it aids comparison to prior numbers.

We use two granularities, and state the rule once so they are not conflated: *recall is measured over gold spans*, and *precision and F1 are measured over posts*. Because our diagnosis is about *which reactions are recovered*, the central quantity is per-stratum span-recall, the fraction of gold reaction spans in a stratum that a prediction matched.

We complement it with item-level precision, recall, and F1 (an “item” is a post), so that the false positives a selector introduces are attributable to the posts that generate them, and we separately report the false-alarm rate on reaction-free posts, defined as the percentage of reaction-free posts on which the model predicts at least one span. It is a rate over posts, not over spans, so a post drawing several spurious spans counts once. The two granularities give deliberately different numbers for the same comparison. A span-recall gain (implicit recall 44.6 to 51.4 for the minority guarantee over BM25, Figure 1) is larger than the matching item-level F1 gain (+2.72, Table 5), because F1 also charges the added false positives. We read span-recall to locate where reactions are missed and item-level P/R/F1 to price the precision cost.

Because a complex model can amplify rather than correct a weak underlying statistical design in health data (Rocchetti 2026), we state and apply a consistent inferential framework appropriate to each analysis. For metric estimates and paired differences we report 2.5/97.5 percentile bootstrap intervals from $B = 5000$ resamples of test items; comparisons against the fine-tuned encoders additionally resample the five training seeds. For these paired-bootstrap comparisons the two-sided p -value is twice the smaller empirical tail probability of the same bootstrap distribution that produces the interval beside it (Efron and Tibshirani 1993). We report two-sided values because a selection method can help or harm and we observe both signs. Two analyses use their own standard machinery rather than the bootstrap: the logistic-regression odds ratios of Section 5.6 carry model-based intervals and p -values, and the cross-corpus homogeneity test of Section 6.2 uses Cochran’s Q against its chi-square reference. For the four per-model tests treated as one comparison family (Table 5) we also report Holm-adjusted p -values; the selector-versus-selector contrasts of Table 4 are exploratory and are not adjusted, which we note where they are used.

3.2 Corpora

Table 1 summarizes the corpora. We use CADEC (Karimi et al. 2015) and PsyTAR (Zolnoori et al. 2019) as our primary study corpora, and add two further corpora, CADECv2 (Dai et al. 2024) and PHEE (Sun et al. 2022), in the cross-corpus robustness analysis of Section 6; those two are described there.¹

The two primary corpora differ in exactly the way our study needs. CADEC posts are reaction-dense and explicit-leaning. Nearly nine posts in ten contain a reaction and, under our validated register labels (Section 5), only about 12% of gold spans are implicit. PsyTAR is the opposite, with reaction-bearing sentences in the minority (about 29% positive) and about 30% of gold spans implicit. The minority guarantee in our selector can only “bite” where reaction-bearing examples are scarce, so PsyTAR is where we expect the method to matter and CADEC is a useful null. Granularity also differs (CADEC is post-level, PsyTAR is sentence-level); we keep per-corpus analyses separate and never pool spans across corpora. Where Section 6.2 reports a pooled figure, it combines the four per-corpus estimates, not their spans.

To make the register distinction concrete, Table 2 gives representative gold reaction spans of each type. We define the two registers by a linguistic criterion, not

¹We do not use the SMM4H adverse-drug-reaction tweet corpus (Magge et al. 2021); its tweet text is access-restricted under platform terms of service, and none of our claims concerns micro-blog text.

Table 1 Corpora used in this study. The roles in the last column summarize the register mix and positive rate discussed in the text. The CADEC test split follows the benchmark we reproduce (Elbiach et al. 2025) and the PsyTAR split is the repository split.

Corpus	Register	Docs	ADR spans	Test items	Role
CADEC	patient forum posts	1250	6318	188	explicit-dense, positive-abundant
PsyTAR	psychiatric drug reviews	5041 sent.	3250	1191	implicit-rich, positive-scarce

Both corpora are previously published and publicly available. PsyTAR counts are sentence-level.

by whether any extractor can match them. An *explicit* reaction denotes a symptom directly, through a clinical or lay term for the condition itself (“nausea”, “insomnia”). An *implicit* reaction conveys the same clinical meaning indirectly, through figurative, narrative, or circumlocutory wording that names no symptom (“not as sharp mentally”, “disconnected from my body”). The distinction is between direct naming and indirect or inferential description. It is a judgment about the wording of the gold span alone, made without reference to any model’s output, so the label of a span does not depend on whether that span was recovered. We use “implicit” as the umbrella term throughout. It subsumes colloquial, figurative, and narrative descriptions, which are overlapping and which we do not separate further here. Section 3.5 explains how we assign the label to every gold span.

Table 2 Representative gold reaction spans by register. Explicit spans are clinically named symptoms; implicit spans convey a reaction in everyday or figurative language. All examples are verbatim gold spans drawn from the labelled test sets and carry the stated register under the validated labeller of Section 3.5, so each can be checked against the released span labels.

Register	Example gold reaction spans
Explicit (clinically named)	“nausea”, “insomnia”, “dizziness”, “weight gain”, “headaches”
Implicit (colloquial / figurative)	“not as sharp mentally”, “feel like you are constantly wearing heavy socks and gloves” (CADEC); “zoned out”, “Zombie effect”, “disconnected from my body”, “Lack of feeling for anything” (PsyTAR)

3.3 Models

Our primary model is GPT-4o-mini, matching the benchmark we reproduce. To show the diagnosis is a property of the task and not of one model, we replicate the key stratified comparison on three additional models spanning scale and family: GPT-4o (OpenAI 2024), Llama-3.3-70B-Instruct (Grattafiori et al. 2024), and Qwen-2.5-72B-Instruct (Qwen Team 2024). All models are queried at temperature 0 through

an OpenAI-compatible API, with five in-prompt examples ($k = 5$) and an identical prompt template across selectors. A small fraction of API calls (under 0.3% on the largest model, from the run logs) returned empty after retries and are scored as no prediction.

3.4 The Demonstration Selector (the Instrument)

For a given test post, a selector chooses the $k = 5$ examples placed in the prompt. A demonstration is, by definition, a labelled input-output pair (few-shot prompting cannot work otherwise), so the demonstration pool already carries gold answers; our selector uses nothing beyond those. It needs no test-time labels, no annotation beyond the demonstrations, no trained retriever, and no gradient training, and it runs on a CPU. We compare the following strategies:

- **random**: examples drawn uniformly from the training pool. This reproduces the benchmark’s few-shot setting (Elbiach et al. 2025) and is the reference point.
- **bm25**: the top- k training posts by BM25 lexical similarity to the test post (Robertson and Zaragoza 2009).
- **sbert**: the top- k training posts by dense sentence-embedding cosine similarity (Reimers and Gurevych 2019).
- **cluster**: a diversity-driven selector that samples from k-means centroids of the training embeddings, the fixed strategy used by a recent ADR-detection baseline (Li and Lin 2026).
- **ours_pos** (retrieval $\cup$ minority guarantee): BM25 retrieval, but reserving several prompt slots for reaction-bearing demonstrations. Reaction-bearing examples are read off the gold answers already present in the demonstration pool (a demonstration whose gold reaction list is non-empty), so the guarantee adds no annotation beyond what few-shot prompting already needs. This is the instrument.
- **ours_full**: ours_pos plus a hard-example component that, within the top- $M=50$ retrieved posts, prefers a positive example whose reaction wording is rare (high mean inverse document frequency). We include it only to ablate it.

Implementation. BM25 uses Okapi weighting over lowercased alphanumeric tokens (Robertson and Zaragoza 2009); dense retrieval and clustering use all-MiniLM-L6-v2 sentence embeddings (Reimers and Gurevych 2019), with k-means run for $\min(k, n)$ clusters (the example nearest each centroid is taken, giving a fixed, query-independent demonstration set). The minority guarantee composes a prompt by walking the BM25 ranking and adding the highest-ranked reaction-bearing demonstrations until $n_{\text{pos}} = 3$ of the $k = 5$ slots are reaction-bearing, then filling the remaining slots with the highest-ranked demonstrations regardless of class; the chosen examples are finally ordered so the most similar one sits nearest the query. The hard variant additionally reserves $n_{\text{hard}} = 1$ slot for the hardest (rarest-wording) positive found within the top $M = 50$ retrieved posts, a relevance gate that keeps the hard example on topic. All selection is deterministic given the seed. The positive quota is fixed at $n_{\text{pos}} = 3$ on every corpus; the selector never measures the positive rate, so its corpus-dependent benefit reflects where reaction-bearing examples are scarce rather than any adaptation to that scarcity.

Algorithm 1 states the minority-guarantee selector (ours_pos) formally. It is entirely lexical and reads only the demonstration pool’s existing gold answers; there is no training and no test-time labelling.

Algorithm 1: Minority-aware demonstration selection (ours_pos).

Input: test post q ; demonstration pool D with gold answers $\text{gold}(\cdot)$; shots k ;
positive quota n_{pos}
Output: ordered list S of k demonstrations
 $R \leftarrow$ the demonstrations in D sorted by BM25 similarity to q (descending);
 $P \leftarrow \{d \in D : \text{gold}(d) \neq \emptyset\}$; // reaction-bearing demos, from
existing labels
 $S \leftarrow []$;
// reserve up to n_{pos} slots for the top-ranked reaction-bearing
demos
foreach $d \in R$ **do**
 if $|S \cap P| \geq n_{\text{pos}}$ **or** $|S| \geq k$ **then break**;
 if $d \in P$ **and** $d \notin S$ **then** append d to S ;
// fill remaining slots with the highest-ranked demos regardless
of class
foreach $d \in R$ **do**
 if $|S| \geq k$ **then break**;
 if $d \notin S$ **then** append d to S ;
sort S by BM25 rank so the most similar demonstration sits nearest q ;
return S ;

The design is deliberately minimal so that the diagnosis is not confounded by a heavy method. The retrieval component tests the “examples should look like the input” hypothesis; the minority-guarantee component tests the “scarce reaction-bearing examples must be present” hypothesis; the hard-example component tests whether up-weighting rare wording helps. As Section 4.3 shows, the hard-example component adds nothing where the minority guarantee acts (on PsyTAR the two agree to within half an implicit F1 point) and costs aggregate F1 where it does not (on CADEC, 59.23 against 62.06), so we report the minority guarantee as the method and treat hard mining as a negative result.

3.5 Register Labelling: Explicit versus Implicit

The diagnosis needs every gold span tagged as explicit or implicit. We first tried a lexicon heuristic and found it unreliable (Section 5); the labels used in all headline results come from a language-model labeller (GPT-4.1-mini) that classifies each gold span in its post context, validated against human annotation. Section 5 reports that validation in full. The labeller is applied uniformly to gold spans across all selectors and all four corpora; it never sees model predictions, so it cannot advantage any selector.

3.6 Fine-Tuned Upper Bound

As the supervised reference line we fine-tune two biomedical encoders, PubMedBERT-base (Gu et al. 2021) and Bio_ClinicalBERT (Alsentzer et al. 2019), as BIO token-classifiers on PsyTAR (train 3850 sentences, test 1191). Each architecture is trained for five epochs (learning rate 3×10^{-5} , batch size 16) over five random seeds (42 to 46), and we report the resulting mean $\pm$ sd band. We verified that the BIO encoding is faithful on PsyTAR (every gold span recoverable from the tag sequence; zero set mismatch on the test split), so the encoders are not handicapped by the label encoding. We do *not* fine-tune on CADEC. Its discontinuous and overlapping entities are only about 82% recoverable under BIO, which would make a CADEC fine-tuned ceiling unfair; where a CADEC fine-tuned number is needed we cite the benchmark’s reported value (strict F1 71.69) (Elbiach et al. 2025). The fine-tuned models are scored with the same scorer and stratified with the same register labels as the in-context runs, so their recall is directly comparable. We call this line an upper bound in the usual sense of a supervised ceiling for the task, not as a bound on what is achievable. On explicit spans the training-free selector in fact matches it (Section 4.5).

4 Results

4.1 Pipeline Validity: Reproducing the Anchor Baseline

Before drawing any conclusion from our pipeline we confirm it reproduces the published baseline. On the CADEC test split (188 posts), our random-exemplar few-shot run reaches strict F1 56.32 against the benchmark’s reported 55.55 (Elbiach et al. 2025), a 0.8-point difference (Table 3). This parity run and the **random** row of Table 4 (56.73) are two independent draws of the same nominal condition, made by different samplers, and they differ on 26 of 188 posts. The 0.4-point spread between them illustrates how much demonstration choice alone can move aggregate F1, though two draws cannot quantify it. That spread is small against the sampling noise in these estimates, roughly a twentieth of the width of the **random** bootstrap interval in Table 4. The parity run draws its five examples with the reproduction script’s own sampler, and the Table 4 row with the selector’s **random** strategy (seed 42, per-item offset), which is why the two differ.

Relaxed F1 is higher (72.72 versus 65.57), but relaxed matching depends on the sentence embedder, and ours differs from the anchor’s, so the two relaxed numbers are not directly comparable. We take strict F1 as the parity target and our primary metric. The parity check licenses the rest of the study, so the differences we report below reflect selection strategy and stratification.

4.2 Aggregate F1: Retrieval Helps, the Minority Guarantee Looks Flat

Table 4 reports aggregate strict F1 for every selector on both primary corpora. Two patterns hold across corpora. First, retrieval outperforms random selection. BM25 gains about +4 F1 on both CADEC and PsyTAR ($p < 0.001$), and dense retrieval gains a similar amount on CADEC, confirming the entity-sensitive retrieval finding

Table 3 Pipeline parity on CADEC (GPT-4o-mini, 5-shot random exemplars, 188 test posts). Strict F1 matches the published benchmark; relaxed F1 is embedder-dependent.

Setting	Strict F1	Relaxed F1
Benchmark (reported) (Elbiach et al. 2025)	55.55	65.57
Our reproduction	56.32	72.72

from prior biomedical work (Sun et al. 2024). Second, the diversity-only selector is unstable. It performs *worse* on CADEC (-4.18 versus random) yet helps on PsyTAR, a corpus-dependent sign flip we return to in Section 7.

The result that motivates the rest of the paper is what the minority guarantee does in aggregate, almost nothing. On CADEC, the guarantee edges BM25 by $+1.23$, which is not significant two-sided ($p = 0.189$), unsurprising because CADEC is positive-abundant so the guarantee is nearly a no-op. On PsyTAR, where it should matter most, it shows no significant aggregate-F1 difference from BM25 ($+0.06$, not significant; this is a failure to detect a difference, not evidence of equality). The hard-mining variant is likewise flat-to-negative. Read only through aggregate F1, one would conclude the minority guarantee is useless. Sections 4.3 and 4.4 show why that conclusion is an averaging artifact.

Table 4 Aggregate strict F1 by selector (GPT-4o-mini, 5-shot, $B = 5000$ bootstrap). Deltas are paired bootstrap differences, with their paired-difference intervals in Appendix F; these selector-versus-selector contrasts are exploratory and are not Holm-adjusted. Significance is two-sided (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Retrieval outperforms random on both CADEC and PsyTAR in these exploratory comparisons; the minority guarantee (ours_pos) is flat in aggregate on PsyTAR, the corpus where it should help most.

Selector	Strict F1 [95% CI]	Δ vs. random	Δ vs. bm25
<i>CADEC</i> (188 posts; explicit-dense, positive-abundant)			
random	56.73 [52.4, 61.0]	—	-4.11
cluster	52.55 [48.2, 57.0]	-4.18^{**} (worse)	-8.29^{***}
sbert	59.95 [56.1, 64.0]	$+3.23^{**}$	—
bm25	60.84 [56.8, 64.7]	$+4.11^{***}$	—
ours_pos	62.06 [58.0, 66.2]	$+5.33^{***}$	$+1.23$ ($p=0.189$)
ours_full	59.23 [55.1, 63.3]	$+2.51^*$	-1.60 ($p=0.109$)
<i>PsyTAR</i> (1191 sentences; implicit-rich, positive-scarce)			
random	55.31 [50.3, 59.9]	—	-4.43
bm25	59.74 [55.4, 63.8]	$+4.43^{***}$	—
cluster	59.92 [55.3, 64.2]	$+4.61^{***}$	$+0.18$ (n.s.)
ours_pos	59.80 [55.7, 63.8]	$+4.49^{**}$	$+0.06$ (n.s.)
ours_full	60.04 [55.9, 64.0]	$+4.73^{***}$	$+0.30$ (n.s.)

4.3 Stratified Diagnosis: The Gap Is Implicit-Shaped

Splitting recall by register is where the picture changes. On PsyTAR under the validated labels, the baseline recovers explicit reactions far more often than implicit ones, with explicit span-recall about 72% versus implicit about 34%, a 38-point gap (Figure 1, left panel). Implicit reactions are roughly twice as hard to recover as explicit ones, and this is what the aggregate F1 was collapsing across. A single test item shows what the deficit looks like in practice. Given a post whose gold reactions are “shaky”, “weak” and “felt like I was going to faint if I stood at all”, the model returns the two explicit spans exactly and a truncated version of the third, which strict matching scores as a miss (Appendix A gives the full prompt and output). This gap is measured on the labeller’s stratification, but it is not an artifact of that labeller. Because label noise blurs the two strata toward each other, any disagreement between the labeller and human annotators (agreement $\kappa = 0.875$; Section 5.1) would attenuate the observed gap rather than inflate it *if* the misclassification were non-differential with respect to whether a span is recovered. We have not established that, so we report the 38-point figure as an estimate under measurement error rather than as a bound. The same ordering holds on CADEC, with a smaller estimated gap; Section 6.2 tests whether the gaps genuinely differ across corpora and does not detect a difference.

Against that backdrop the minority guarantee does exactly what aggregate F1 hid. Table 5 reports the item-level implicit F1 difference between the minority guarantee and BM25 across all four models. The gain is positive on every one of the four models (+1.95 to +2.72) and precision-preserving, with implicit precision essentially unchanged on the primary model (74.0 to 73.8) while implicit recall rises. Rather than lean on any single model, we run a pooled analysis across them. A paired item bootstrap of the mean implicit-F1 gain across the four models, resampling posts so that the shared-item dependence between models is respected, gives a combined gain of +2.35 (95% CI [+0.72, +4.14], two-sided $p = 0.003$). All pooled figures in this paper come from this single implementation, so they are directly comparable with the sensitivity analysis of Section 5.4.

The pooled effect is therefore significant, even though per model it is small enough that the individual two-sided paired-bootstrap p -values (0.040 to 0.138) do not clear Holm correction (Table 5); the same-signed gain also recurs under a non-GPT labeller (Section 5.5). This is the profile of an instrument, a small but reliable effect that exposes the diagnosis without itself topping a leaderboard.

On GPT-4o, the model with the least explicit headroom, the gain is larger on the implicit stratum than the explicit one (+2.6 against +1.3 F1); on the other three, which still have explicit headroom, it helps both. The implicit gain is the consistent one, present on every model, which is the thesis in one line, a training-free intervention helps where in-context learning is weakest. On CADEC the same comparison is not significant (only about 12% of spans are implicit and positives are abundant), so the method’s benefit is not universal. Section 6.3 identifies what it tracks, namely how scarce reaction-bearing examples are in the demonstration pool, and shows that the guarantee barely alters the prompt on the three corpora where they are plentiful. The hard-mining variant adds nothing on top of the guarantee (implicit item-level F1 70.2 against 70.7), confirming that the entire effect is the minority guarantee.

Table 5 Stratified item-level implicit F1: ours_pos versus bm25, by model (PsyTAR, 5-shot, validated LLM register labels, $B = 5000$ paired bootstrap). p is two-sided; the guarantee can add or remove reaction-bearing examples from a prompt and we observe both signs across corpora, so a one-sided test is not justified. p_{Holm} is Holm-adjusted across the four-model family. The gain is positive on every model, and precision-preserving on every model (Appendix B); per-model evidence is weak and does not clear Holm correction, so the evidence is the consistent sign and size across the four models (and under a non-GPT labeller), not any single test.

Model	Δ implicit F1	95% CI	p (2-sided)	p_{Holm}
GPT-4o-mini	+2.72	[+0.11, +5.69]	0.040	0.158
GPT-4o	+2.53	[−0.51, +5.72]	0.107	0.320
Llama-3.3-70B	+1.95	[−0.55, +4.64]	0.138	0.320
Qwen-2.5-72B	+2.19	[−0.67, +5.49]	0.138	0.320

Baseline PsyTAR span-recall for reference: explicit $\approx 72\%$, implicit $\approx 34\%$ (38-point gap). Pooled test across the four models (paired item bootstrap of the mean gain): +2.35 (95% CI [+0.72, +4.14], two-sided $p = 0.003$).

Figure 1, the central figure of the paper, shows per-stratum recall across selectors, with the fine-tuned upper bound drawn as a reference line (Section 4.5). The explicit panel is high and its selectors sit close together; the implicit panel is low, and it is where the selectors, and the fine-tuned line, separate.

4.4 Operating Point: Pareto Gain on Positives, False-Alarm Cost on Negatives

Why is a clear per-stratum benefit invisible in aggregate F1? Because the cost lands on a different population than the benefit. Table 6 decomposes it. On posts that contain a reaction, the minority guarantee improves recall on both strata, and its precision effect differs by stratum. On explicit spans precision and recall both rise (a Pareto move, meaning neither is traded for the other, placing the guarantee above BM25’s precision-recall frontier, 74.7/75.9 to 77.0/79.7), and on implicit spans recall rises at essentially unchanged precision (62.9 to 67.9 recall, 74.0 to 73.8 precision). Either way it is not trading precision away on real-ADR posts, so the recall it buys there is close to free. The cost is entirely on reaction-free posts, where the false-alarm rate rises from 20.1% to 25.8% (+5.7 points). That extra sensitivity on true positives and extra false alarms on true negatives roughly cancel in a corpus-wide F1, which is precisely why aggregate F1 is flat.

The net effect is a sensitivity-oriented operating-point shift that catches more reactions at the cost of more false alarms on clean posts, a trade that is defensible for a pharmacovigilance triage stage where recall is at a premium, but that must be shown as a trade rather than a costless gain. A characterization of the added false positives (Section 5.7) finds them predominantly textually grounded, with hallucinations rare, but disproportionately of the indication-confusion type, so the trade is real and we quantify its composition there.

When is this trade worth taking? The minority guarantee is the right operating point for a high-recall first-pass filter, the stage that feeds a human reviewer or a downstream classifier, and it earns its keep precisely where reaction-bearing posts are

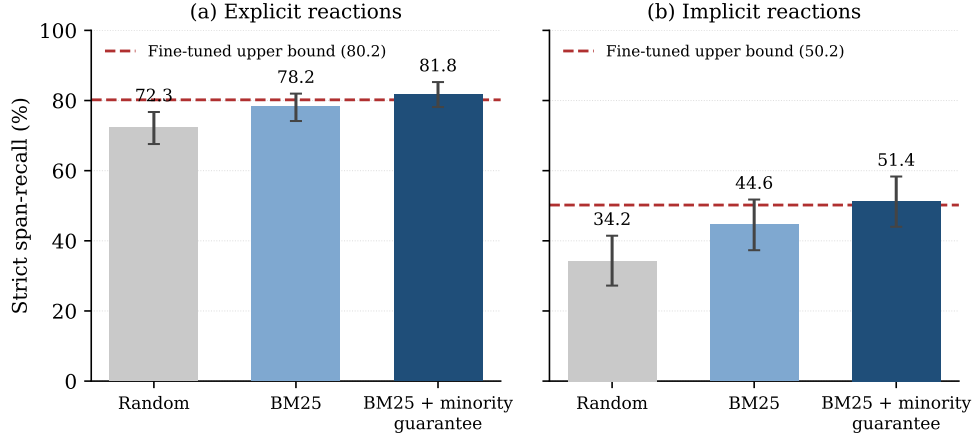


Fig. 1 Per-stratum strict span-recall on PsyTAR (GPT-4o-mini, 5-shot; 528 explicit and 222 implicit gold spans under the validated register labels), with 95% bootstrap confidence intervals and the PubMedBERT fine-tuned upper bound (five-seed mean) as a horizontal reference line. **(a)** Explicit reactions: recall is high throughout (72.3 to 81.8) and the selectors are closer together than in **(b)**; only the minority guarantee rises above the fine-tuned line. **(b)** Implicit reactions: recall is far lower; the selectors separate cleanly (random 34.2, BM25 44.6, minority guarantee 51.4), and the training-free minority guarantee reaches the fine-tuned band (mean 50.2) using only the demonstrations’ own labels. The bar sits nominally above that line, but the two are not distinguishable in a paired test on the same items (Section 4.5), so the panel should be read as the selector reaching the band, not beating it. Even that supervised upper bound leaves a large implicit deficit relative to the explicit panel.

Table 6 Operating-point decomposition on PsyTAR (GPT-4o-mini, item-level strict P/R under the validated LLM register labels). On explicit spans the minority guarantee improves both precision and recall (a Pareto move); on implicit spans it improves recall at essentially unchanged precision; the only cost is a higher false-alarm rate on reaction-free posts.

Population	bm25 P	bm25 R	ours_pos P	ours_pos R
positive posts, explicit spans	74.7	75.9	77.0	79.7
positive posts, implicit spans	74.0	62.9	73.8	67.9
negative (no-ADR) posts, false-alarm rate	bm25 20.1% → ours_pos 25.8% (+5.7 pp)			

Positive-post cells are item-level strict precision and recall on the same validated register labels as Table 5 (341 reaction-bearing posts: 159 explicit-only, 182 implicit-bearing); the false-alarm rate is over 850 reaction-free posts and is label-independent.

scarce enough that random exemplars miss them, lifting post-level detection sensitivity, the percentage of reaction-bearing posts on which at least one span is predicted, from 86.5% to 92.7%. It is not a free improvement. It also lifts the false-alarm rate on reaction-free posts from 20.1% to 25.8%, so at low positive prevalence (10 to 30%) it gives back a few points of post-level precision, which combines the two rates as $\pi s / (\pi s + (1 - \pi)f)$ for prevalence π , sensitivity s and false-alarm rate f (for instance 28.6% against BM25’s 32.3% at $\pi = 0.10$). Where precision is paramount or positives

are very rare, BM25 alone is the better choice. The point is that this is a characterized operating-point choice with a clear regime, not a uniform win or loss.

4.5 Fine-Tuned Upper Bound: The Residual Survives Fine-Tuning

The decisive question is whether the implicit deficit is an artifact of in-context prompting that fine-tuning would erase. It is not. We fine-tuned two independently pretrained encoder taggers directly on the PsyTAR gold labels, PubMedBERT and Bio_ClinicalBERT, with five random seeds each, and stratified every run (Table 7). The seed-42 PubMedBERT aggregate strict F1 (71.15) matches the fine-tuned ballpark from the benchmark (Elbiach et al. 2025), so these are fair upper bounds. Yet across all ten runs the explicit-versus-implicit span-recall gap is 30 to 32 points (PubMedBERT 30.0 ± 1.9 , Bio_ClinicalBERT 32.2 ± 2.4), nearly as wide as the 38-point gap of the in-context baseline and roughly fifteen times the seed-to-seed variation. Training on the labels narrows the gap by under 10 points, and a second, differently pretrained encoder reproduces the same shape, so the residual is a property of the task, not of one model.

Two comparisons make the point sharp. First, we compare the selector against each fine-tuned encoder *on the same test items*, with a hierarchical paired bootstrap of the difference in strict span recall that resamples both test items and the five training seeds, so the interval reflects training-seed variability rather than conditioning on the particular runs we trained. Comparing two separately estimated intervals for overlap would not test this. Against PubMedBERT the selector is +1.17 points on implicit recall (95% CI $[-5.60, +7.75]$, two-sided $p = 0.73$) and +1.59 points on explicit recall ($[-2.43, +5.60]$, $p = 0.44$); against Bio_ClinicalBERT it is +4.32 ($[-2.38, +11.18]$, $p = 0.22$) and +2.58 ($[-1.42, +6.65]$, $p = 0.22$). No difference is detectable in either register, and the point estimate favours the training-free selector in all four comparisons.

Absence of a detectable difference is not proof of equivalence, and with these interval widths the data exclude a fine-tuned recall advantage larger than about 5.6 points on implicit spans and 2.4 points on explicit spans, but cannot rule out a smaller one. The selector reaches this without fine-tuning, without a GPU, and without annotation beyond the demonstrations.

Second, where fine-tuning genuinely wins is elsewhere. It lifts both strata off the in-context floor (implicit from 34 to about 50, explicit from 72 to about 80) and slashes the false-alarm rate on reaction-free posts (from 20 to 26% in context down to about 7.5%). Fine-tuning’s demonstrated advantage over the selector therefore lies in precision rather than recall. The selector’s explicit span recall (81.8%) even slightly exceeds PubMedBERT’s (80.2 ± 0.5). Fine-tuning is overwhelmingly better on reaction-free posts, and modestly better on reaction-bearing ones, where its implicit item-level F1 is 73.2 ± 0.6 against the selector’s 70.7. False-alarm suppression is the larger of the two and the principal demonstrated source of the aggregate difference, though not the only one.

The implicit deficit survives the fine-tuning boundary across seeds and architectures, which promotes the diagnosis from “in-context learning is bad at implicit

reactions” to “implicit reactions are harder to detect even for a supervised upper bound,” within the encoder setups we evaluated.

Table 7 Fine-tuned upper-bound band on PsyTAR (strict span-recall over 528 explicit / 222 implicit gold spans; validated register labels; five random seeds per architecture, mean $\pm$ sd). Two independently pretrained encoders both leave a 30 to 32 point explicit-versus-implicit recall gap, far larger than the seed variation, so the residual is not a PubMedBERT artifact; it persists across both architectures and all ten training runs we evaluated.

Architecture (5 seeds)	Explicit rec.	Implicit rec.	Gap (E-I)	Implicit item F1	Neg. FA%
PubMedBERT	80.2 $\pm$ 0.5	50.2 $\pm$ 2.0	30.0 $\pm$ 1.9	73.2 $\pm$ 0.6	7.4 $\pm$ 0.6
Bio_ClinicalBERT	79.2 $\pm$ 0.9	47.0 $\pm$ 1.6	32.2 $\pm$ 2.4	70.9 $\pm$ 1.2	7.6 $\pm$ 0.8

In-context baseline gap for reference: $72.3 - 34.2 = 38.1$ pt. Seed-42 PubMedBERT aggregate strict P/R/F1 = 70.31 / 72.0 / 71.15 (cf. benchmark fine-tuned $\approx$ 71.7 (Elbiach et al. 2025)). The best training-free selector reaches implicit recall 51.4 [44.0, 58.3], overlapping the top of this band; the in-context false-alarm rate is 20 to 26% versus about 7.5% here.

5 Validity and Robustness

A diagnosis that lives in a subgroup and a non-primary metric invites the charge of cherry-picking. The checks below address it. The register labels are human-validated and reproduce under a labeller from a different model family, the advantage is not leakage, it does not depend on borderline spans, and it is not a proxy for length or rarity.

5.1 Register-Labeller Validation

Two annotators, the two authors, independently labelled a balanced sample of 180 gold spans (with sentence context) as explicit or implicit following the written codebook (Appendix D). They labelled blind to the model predictions, to the heuristic label, and to the LLM label, though not blind to corpus or to the study hypothesis, a limitation we return to in Section 8. Raw agreement was 89.4%. The full confusion is 144 spans both called explicit, 17 both called implicit, 19 the more liberal annotator called implicit while the other called explicit, and 0 the reverse.

Two things follow. First, despite the balanced 90/90 sampling on the heuristic label, the annotators’ realized split is about 80% explicit (annotator A called 144 of 180 explicit), so one class is in fact rare, which is exactly the condition that deflates Cohen’s κ (the “ κ paradox”). Cohen’s κ was 0.589 (95% CI [0.42, 0.74]), but the prevalence-adjusted PABAK was 0.789 [0.69, 0.88] and Gwet’s AC1 was 0.859 [0.79, 0.92] (Cohen 1960; Byrt et al. 1993; Gwet 2008), both indicating good agreement. Second, the disagreements are one-directional (all 19 in the same direction), so they are adjudicable and, as Section 5.3 shows, immaterial to the result. Table 8 summarizes.

Crucially, this human study *invalidated* our first labeller. A lexicon heuristic agreed with the humans at only κ 0.19 to 0.40, over-counting implicit reactions two- to three-fold and missing standard symptom terms; every number in this paper uses instead a

language-model labeller (GPT-4.1-mini) that agrees with the human consensus at κ 0.875, chosen over GPT-4o (0.73) and Qwen (0.77). Replacing the heuristic with the validated labeller did not weaken the finding; it sharpened it (the baseline implicit-explicit gap widened from about 28 to 38 points, because the heuristic had been mislabelling easy spans as implicit). The stratification of the full test set is produced by a language model, not a full human gold standard; we treat its agreement with the 180-span human consensus as validation on a sample, and lean on the boundary-sensitivity and cross-family checks (Sections 5.3 and 5.5) rather than on the labeller’s authority.

Table 8 Register-labeller validation. Human inter-annotator agreement is good once prevalence is accounted for; the LLM labeller used for all results agrees with the human consensus at κ 0.875, while the discarded lexicon heuristic did not.

Comparison	Agreement	Notes
Human vs. human (180 spans)	raw 89.4%; κ 0.589; PABAK 0.789; AC1 0.859	disagreements one-directional
LLM labeller (GPT-4.1-mini) vs. human consensus	κ 0.875	chosen labeller
GPT-4o vs. human consensus	κ 0.73	not chosen
Qwen vs. human consensus	κ 0.77	not chosen
Lexicon heuristic vs. human	κ 0.19 to 0.40	invalidated; discarded

5.2 Leakage Audit

A retrieval-based method could “win” merely by copying a near-duplicate of the test post (with its answer) into the prompt. We audited train/test near-duplication with dense cosine similarity above 0.9 or token Jaccard above 0.8. Near-duplication is low (CADEC 1.1%, PsyTAR 3.1%) and answer-leaking near-duplicates are negligible (CADEC 0.0%, PsyTAR 0.3%, four items). Removing the near-duplicates barely moves the headline. The de-leaked PsyTAR implicit advantage of the minority guarantee over BM25 is +2.76 (95% CI [+0.17, +5.62], two-sided $p = 0.038$), essentially identical to the full-data +2.72 ([+0.12, +5.64], $p = 0.039$). The advantage is not contamination.

5.3 Boundary Sensitivity

Some spans are genuinely borderline between explicit and implicit. Two independently validated labellers (GPT-4.1-mini and GPT-4o) agree on 92.1% of 750 PsyTAR spans, leaving 59 borderline spans. We recomputed the headline under four labellings, each labeller alone plus the borderline spans forced entirely to implicit or entirely to explicit. Across all four, the explicit-implicit recall gap stays in the 36.8 to 39.9 point band and the guarantee’s implicit F1 advantage over BM25 stays between +2.7 and +3.6, all two-sided $p \leq 0.039$. The finding is independent of the 59 borderline spans, so no

manual adjudication of them is needed. Moreover the human disagreements are one-directional, and every alternative labelling above yields a larger advantage than the one we report, so the reported figure is the conservative choice.

5.4 Register-Label Consistency

The check above varies where the register line is drawn. A separate question is whether the released labels are internally consistent with the codebook of Appendix D. We therefore ran reproducible consistency checks over all 750 labelled PsyTAR spans. These are pattern-based rather than a semantic re-reading of every span, so they establish the absence of certain inconsistencies rather than the absence of all of them. Of the 126 spans matched by the audit’s clinical-vocabulary pattern, 124 are labelled explicit as rule (1) requires. The insomnia lay-synonym check matched 12 spans (“can’t sleep”, “cannot sleep at night”, “trouble sleeping”); all 12 are labelled explicit under rule (1), while sleep-related descriptions that require inference (“slept through life”, “spend most of my spare time in bed sleeping”) are labelled implicit.

The two exceptions are “Very apathetic” and “mild apathy”. Both contain a clinical term and should be explicit under rule (1), so we count them as labelling errors rather than defensible context-dependence, and we release a corrected label file alongside the original. Separately, five normalized span strings, covering 21 of 750 spans (2.8%), receive both labels in different posts: “vivid dreams”, “anger”, “gained weight”, “sexual side effects” and “brain fog”. Because the labeller sees each span inside its host post, some context-dependence is by design, but these five are not distinctions we would defend.

We therefore re-ran the pooled cross-model test under four alternative labellings, all using the same implementation as Section 4.3 so that they are directly comparable with it and with one another. Correcting only the two apathy errors gives +2.41 ([+0.74, +4.24]). Forcing all 21 ambiguous spans to explicit gives +2.54 ([+0.81, +4.50]) and to implicit +2.24 ([+0.66, +4.00]); correcting the apathy errors as well gives +2.61 ([+0.83, +4.57]). The released labels give +2.35, and two-sided $p \leq 0.004$ throughout. The conclusion is unchanged across all five conditions.

The primary analyses throughout the paper use the original released labels, including the two spans we identify as errors. The original label set was frozen before any results were examined. To preserve that analysis and to avoid outcome-informed relabelling, we retain it as primary while releasing the corrected labels beside it and reporting their isolated effect, which is the second condition above. The audit, both label files and the sensitivity analysis are released with the code, so every count in this section can be reproduced.

5.5 Cross-Family Labeller Robustness

The boundary check above varies the label within the GPT family; a stronger test is whether the register axis is itself an artifact of using a GPT model to draw it, given that the primary extractor is also a GPT model. We therefore re-derived the entire stratification with a labeller from a different model family, Qwen-2.5-72B, which was validated against the human consensus at κ 0.77.

Under the Qwen labels the split shifts slightly (25.2% of spans implicit, versus 29.6% under GPT-4.1-mini), but the diagnosis is unchanged. Baseline explicit span-recall is 71.5% against implicit 30.2%, a 41-point gap, even wider than the 38-point gap under the GPT labeller, and the minority guarantee’s implicit F1 advantage over BM25 holds at the same magnitude, +2.31 (95% CI $[-0.30, +5.02]$, two-sided $p = 0.086$), which like the per-model GPT-labelled tests is not individually significant; the point of the check is that the register axis and the direction of the effect survive a change of labeller family, not that this single comparison is decisive. The implicit deficit is therefore not an artifact of a GPT model drawing the line, and the selector’s benefit points the same way and by a similar amount when a different model family draws it.

5.6 Confound-Matching: Implicit Is Not Length or Rarity in Disguise

Implicit spans tend to be longer and could use rarer words, so the register gap might be a proxy for span length or word rarity. Our primary test is a logistic regression of “recovered” on an implicit indicator, with and without controls for span token length and mean token log-frequency (Table 9). The controls attenuate the implicit effect (odds ratio 0.20 to 0.36 for the in-context baseline, 0.26 to 0.40 for the fine-tuned model), so part of the raw gap is length; but the penalty does not vanish. After controls, an implicit span still has about 0.36 to 0.40 times the odds of recovery of an explicit span of the *same* length and rarity ($p \leq 10^{-6}$), in both the prompt-only baseline and the supervised upper bound. Length is the stronger single control (long spans are genuinely hard) and correlates with implicitness ($r = 0.33$); rarity is nearly irrelevant.

As a descriptive cross-check, the gap also holds within each marginal bin (Figure 2). Explicit recall exceeds implicit recall at every span length and every word-rarity tercile, with per-bin support given in the caption. A finer length-by-rarity grid is consistent where it is populated, but we do not lean on it because several implicit cells are sparse or empty. The narrow claim we defend is that implicit is not a proxy for length or rarity, not that length and rarity are irrelevant; controlling for both attenuates the gap but leaves a large, highly significant implicit penalty.

5.7 False-Positive Characterization

Because the operating-point trade adds false alarms on reaction-free posts, we characterized every one of them on the 850 reaction-free posts, and classified the largest bucket rather than leaving it undifferentiated. Each predicted span on such a post is a false positive. We first bucket it as hallucinated (the span text is absent from the post), negation-flip (present, but a negation cue sits beside it), or asserted (present and unnegated); we then classify each asserted span as indication confusion (it names the condition the drug is taken to treat), a plausible reaction the gold annotation missed, or other. Table 10 reports the result.

Hallucination is rare under both selectors, 8 of 275 false positives for BM25 and 12 of 350 for the minority guarantee, so the added errors are predominantly grounded in the post’s own words rather than invented, though 4 of the 75 added are hallucinations.

Table 9 Confound-matching on PsyTAR (750 spans). Logistic regression of span recovery on an implicit indicator, alone (a) and with span length and word log-frequency controls (b). Odds ratios below 1 mean lower recovery for implicit spans. The implicit penalty attenuates under controls but stays large and highly significant for both the in-context baseline and the fine-tuned upper bound.

Model	(a) implicit alone, OR [95% CI]	(b) + length + log-freq, OR [95% CI]
In-context (random baseline)	0.199 [0.142, 0.279], $p=5 \times 10^{-21}$	0.355 [0.237, 0.531], $p=5 \times 10^{-7}$
Fine-tuned (PubMedBERT)	0.260 [0.185, 0.366], $p=9 \times 10^{-15}$	0.399 [0.276, 0.577], $p=1 \times 10^{-6}$

Controls are z-scored; token.length odds ratio is 0.12 (baseline) / 0.45 (fine-tuned), $p < 10^{-11}$, log-frequency is near 1. A finer length $\times$ frequency grid is directionally consistent but has sparse implicit cells, so we rely on the regression and the marginal bins of Figure 2.

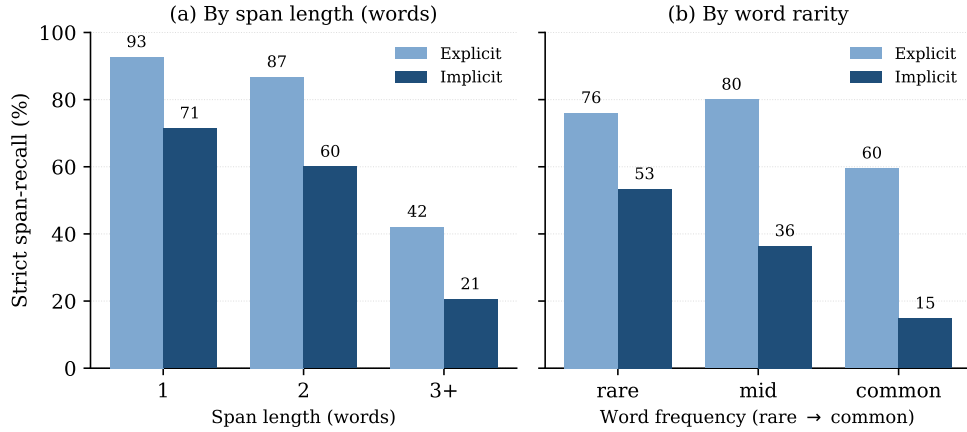


Fig. 2 Explicit versus implicit strict span-recall within matched bins on PsyTAR (in-context random baseline). (a) By span length in words: implicit recall is lower at every length, so the gap is not just that implicit spans are longer. (b) By word rarity (frequency terciles, rare to common): implicit recall is lower at every rarity level. The register gap holds inside every bin, evidence that it is not merely a proxy for span length or word rarity. Bin sizes (explicit / implicit spans): lengths 1, 2, 3+ have 150/21, 188/50, 190/151; rarity terciles rare, mid, common have 175/75, 190/66, 163/81.

The added false alarms are not, however, simply more of the same. Of the 75 false positives the guarantee adds, 37, about half, are indication confusion, against only 9 further plausible reactions. That is a genuine pharmacovigilance error rather than a free pass. Naming the depression for which an antidepressant was prescribed is wrong in a way that an over-read but real symptom is not.

The trade is therefore more specific than a uniform rise in false alarms. The guarantee buys implicit recall on reaction-bearing posts at close to no precision cost, and pays for it on reaction-free posts with errors that are textually grounded but weighted toward indication confusion. For a triage stage with a human reviewer this may remain workable, since indication confusion should be quick for a reviewer to dismiss and is potentially addressable downstream by filtering predicted spans that match a drug's

known indications, neither of which we tested. It is not the costless “the same errors, only more of them” picture, and a deployment without such filtering should weigh the 5.7-point false-alarm rise accordingly.

Two caveats apply to the bucketing itself. The negation rule is a crude window heuristic, so the negation-flip share is an upper bound, and the three-way classification of asserted spans is assigned by the same validated labeller family used for register labels, not human-adjudicated.

Table 10 False-positive composition on the 850 reaction-free PsyTAR posts (GPT-4o-mini, 5-shot), as counts. The minority guarantee adds 75 false positives over BM25, of which about half are indication confusion.

Selector	FPS	Halluc.	Negation	Indication	Plausible	Other
bm25	275	8	31	89	121	26
ours_pos	350	12	41	126	130	41
<i>added</i>	<i>+75</i>	<i>+4</i>	<i>+10</i>	<i>+37</i>	<i>+9</i>	<i>+15</i>

6 Cross-Corpus Robustness

Everything above rests on two corpora, which leaves open whether the register deficit is a property of patient-authored text or an artifact of these particular benchmarks. We therefore extend the analysis in two directions, which answer different questions. To ask whether the diagnosis generalizes *within* the colloquial register, we add CADECv2 (MultiADE) (Dai et al. 2024), an independently constructed patient-forum corpus. To ask what happens under genuine *domain shift*, we add PHEE (Sun et al. 2022), a corpus of biomedical case reports whose language is technical rather than colloquial.

CADECv2 and CADEC share a small number of AskaPatient posts (46 of 709 test documents, 6.5%, matched on case-folded alphanumeric text). We remove every shared document from both splits, leaving 663 test documents with no exact match in CADEC, so this is independent evidence rather than a partial re-test of a corpus we have already used. All CADECv2 numbers below are on that de-duplicated split.

6.1 The Register Mix Is Corpus-Dependent

Table 11 applies the labeller validated on the primary corpora (Section 3.5) to all four corpora. The share of implicit spans varies more than threefold, from 8.1% on case reports to 29.6% on psychiatric drug reviews. The technical corpus sits lowest, below all three patient-authored corpora, with an interval disjoint from those of CADECv2 and PsyTAR and only just touching CADEC’s, which is the expected result, since case reports are written by clinicians in clinical vocabulary, so the colloquial register is largely absent by construction. The patient-authored corpora are not interchangeable either. CADEC and CADECv2 are drawn from the same forum yet have clearly different estimated register mixes (12.0% against 17.9%, with disjoint intervals), consistent with their different drug and annotation coverage.

The same table reports two further baseline characteristics. The positive rate ranges from 28.6% to 95.0%, so the corpora span both the reaction-scarce and the reaction-abundant regime, and median document length ranges from 65 to 417 characters, spanning the sentence-level and document-level regimes; we did not separately quantify textual noise. Any claim that survives across these four is not resting on one class balance or one document granularity.

Table 11 The four corpora under the same validated register labeller. Implicit share is over gold reaction spans in the test split, with a 95% bootstrap interval. CADECv2 is the de-duplicated split described above.

Corpus	Source	Items	Pos.%	Med. chars	Spans	Implicit % [95% CI]
PHEE	case reports	968	91.7	126	951	8.1 [6.4, 9.9]
CADEC	patient forum	188	89.4	388	876	12.0 [9.8, 14.2]
CADECv2	patient forum	663	95.0	417	4568	17.9 [16.8, 19.0]
PsyTAR	drug reviews	1191	28.6	65	750	29.6 [26.3, 32.9]

6.2 The Diagnosis Generalizes, Including Off Social Media

We then measured the quantity the paper is about, per-stratum strict span recall, under one fixed condition on all four corpora (BM25 selection, GPT-4o-mini, five shots), so the corpora rather than the configuration are what differ. Table 12 reports the result with a paired item bootstrap.

The explicit-implicit gap appears in every corpus, at between 26.6 and 33.6 points, with every interval excluding zero. Overlapping intervals would not by themselves establish that the four gaps are the same, so we tested it. Cochran’s Q over the four estimates, weighted by their bootstrap standard errors, is 1.29 on 3 degrees of freedom ($p = 0.73$) with $I^2 = 0\%$. We therefore detect no heterogeneity among the four estimated gaps, and the fixed-effect pooled gap is 31.3 points (95% CI [28.0, 34.7]). With only four corpora, moderate differences between them could still go undetected. Absolute difficulty, by contrast, clearly does vary, with explicit recall ranging from 58.7% to 78.2%, so the corpora genuinely differ in how hard they are while the penalty attached to implicit phrasing does not measurably differ.

The PHEE row is the one we expected to break the pattern, and it did not. Case reports are written by clinicians, and accordingly only 8.1% of their reaction spans are implicit, against 29.6% on PsyTAR. But where implicit spans do occur, they are recovered 30.1 points less often, an estimate within the range observed on the patient-authored corpora.

This distinguishes two claims that are easy to run together. The prevalence of implicit phrasing varies substantially by corpus and source domain, more than three-fold across our four. How *much harder* such reactions are to extract, once present, did not measurably differ across those same four, across class balances from 28.6% to 95.0% positive and document lengths from 65 to 417 characters. The deficit is therefore not an artifact of the two benchmarks the earlier sections used, and its appearance

in PHEE is evidence that it can extend beyond social media. PHEE’s implicit stratum is small (77 spans), which is why its interval is the widest in the table.

Table 12 Strict span recall by register on all four corpora under one fixed condition (BM25, GPT-4o-mini, five shots), with 95% paired item-bootstrap intervals. The gap is present in all four corpora; no heterogeneity in its magnitude is detected, whereas absolute recall varies.

Corpus	Explicit %	Implicit %	Gap (pp)	95% CI
PHEE	58.7	28.6	30.1	[19.2, 40.7]
CADEC	64.7	38.1	26.6	[16.9, 36.4]
CADECv2	70.0	38.1	31.8	[27.6, 36.0]
PsyTAR	78.2	44.6	33.6	[25.5, 41.4]

6.3 The Selector’s Benefit Tracks the Demonstration Pool

The third claim, the value of the minority guarantee, behaves differently across corpora. Rather than infer the reason, we measured it. The guarantee can only matter when BM25’s own top- k already lacks reaction-bearing demonstrations, and that depends on the *demonstration pool*, not on the prevalence of the test split, which is the quantity it would be natural but mistaken to cite here. Table 13 reports, for each corpus, the pool’s positive rate, how many reaction-bearing examples BM25 alone places in the five-shot prompt, and how often the guarantee actually changes which demonstrations are shown.

The mechanism is then visible directly. On PsyTAR the pool is 29.7% positive and BM25 supplies only 1.44 reaction-bearing examples per prompt, below the $n_{\text{pos}} = 3$ the guarantee requires, so it rewrites the demonstration set for 78.0% of items. On the other three corpora the pool is 88 to 94% positive and BM25 already supplies between 4.7 and 4.9, so the guarantee has nothing to add and changes the set for between 0.2% and 1.2% of items. It is nearly inert at the prompt-selection level rather than inactive by coincidence.

On CADECv2, accordingly, its implicit item-level F1 is -0.84 against BM25 (95% CI $[-1.63, -0.06]$). This is an exploratory estimate with an unadjusted interval that excludes zero, so on this corpus the guarantee is slightly negative rather than merely inert, by under one F1 point; implicit span recall is essentially unchanged (38.1% to 38.6%). This matches CADEC, which is also positive-abundant, and contrasts with PsyTAR, the one corpus in which positives are scarce (28.6%) and in which the guarantee helps (Section 4.3).

This corpus dependence is what the proposed mechanism predicts. A selection method that helped equally whether or not the minority it targets was actually rare would point to something other than the stated mechanism driving the gain. It is also consistent with Peng et al. (2024), who report that the value of a demonstration-selection strategy depends on the data it is applied to. The practical reading is that

the guarantee materially changed prompts only where reaction-bearing demonstrations were the minority, which is the realistic pharmacovigilance setting of screening unfiltered streams, and that it did not help on the reaction-dense pools evaluated here.

Table 13 Where the minority guarantee can act. It binds only when BM25 supplies fewer than $n_{\text{pos}} = 3$ reaction-bearing demonstrations, which is a property of the demonstration pool rather than of the test split. Selection is deterministic given pool and query, so these figures need no model calls.

Corpus	Pool	Pool pos.%	BM25 pos./5	Prompts changed %	Test pos.%
PsyTAR	3850	29.7	1.44	78.0	28.6
CADEC	1062	88.4	4.82	0.5	89.4
CADECv2	2678	93.8	4.91	0.2	95.0
PHEE	2898	94.1	4.69	1.2	91.7

7 Discussion

The headline is a methodological one. On this task, a single aggregate F1 collapses two populations with very different behaviour into one number, and reporting only that number hides both the real benefit of a minority-aware selector and the real location of the residual gap. Stratifying by how the target is expressed is a cheap, reusable move, and it is the transferable contribution here, more than the selector itself.

The selector is a modest instrument. Its benefit is real but small, precision-preserving on reaction-bearing posts, and confined to corpora where reaction-bearing examples are scarce in the demonstration pool, which among the four we test means PsyTAR alone (Section 6.3). It buys sensitivity at the price of false alarms on clean posts, a trade that a corpus-wide F1 would hide. That the training-free guarantee has a similar implicit-recall point estimate to a supervised encoder, with no detectable difference between them, is the most striking single comparison. The interval still admits a fine-tuned advantage of up to about 5.6 points on this stratum, so what we can say is that the labels buy no detectable recall here, not that they buy nothing.

The corpus-dependent sign flip of the diversity selector (helpful on PsyTAR, harmful on CADEC) is consistent with the wider finding that the value of an example-selection strategy is dataset- and model-dependent rather than universal (Peng et al. 2024). CADEC’s explicit density makes retrieval-similar examples already informative, so forcing diversity discards useful context, whereas PsyTAR’s heterogeneity rewards it. This is a reminder that “smart selection” is not a single knob with a fixed sign.

The deepest point is how far the residual reaches. It persists in three separate sweeps: across four in-context models, across two fine-tuned encoder architectures with five seeds each, and, under one fixed extraction condition, across all four corpora. The difficulty of implicit reactions is therefore not specific to in-context prompting, nor to any one corpus. These reactions must be inferred from figurative or indirect language rather than matched to a symptom name, and that difficulty holds across every model and paradigm we evaluated. This shows the gap is robust to the models, objectives,

and selection methods we tested, not that no conceivable model or training regime could narrow it. This connects to observations that language models struggle with the nuances of non-standard reaction descriptions in social-media health text (Chandra et al. 2025) and reframes “fine-tuning wins” as a statement about where, not whether. Fine-tuning wins principally by suppressing false alarms. The paired comparisons detect no recall advantage in either register, and the implicit deficit is one both paradigms share.

The deficit also compounds downstream. Implicit mentions are precisely the spans a normalizer struggles to map to a standard terminology such as MedDRA, because non-standard wording has no direct dictionary entry (Yazdani et al. 2023). Surfacing them is a prerequisite and already difficult part of that pipeline, so locating where they are lost is useful in its own right, which is why we treat extraction without normalization as a scope choice rather than a gap.

There are two practical takeaways. For anyone reporting on this task, a single corpus-wide F1 is an unreliable summary. It can hide a real intervention effect (as it hides ours) and it obscures which linguistic phenomena a system actually handles. Reporting recall stratified by register, at a stated operating point, costs one extra table and one register labeller, and it turns an opaque ranking into an interpretable one. For anyone deploying an extractor as a first-pass filter in a pharmacovigilance pipeline, the operating-point framing is the actionable part. The minority guarantee is a cheap way to raise sensitivity to reactions when reaction-bearing posts are scarce, at a quantified and characterizable false-alarm cost that a downstream human or model reviewer would have to absorb.

Which paradigm to deploy follows the resource regime. With labelled spans and compute, a fine-tuned encoder is the stronger extractor overall, with far fewer false alarms and a higher aggregate F1, although the paired comparison of Section 4.5 detects no recall advantage for it in either stratum, and the point estimates there favour the selector; without them, the training-free selector reaches comparable recall at no training cost. But the colloquial gap survives both, so a register-stratified check earns its place under either choice. Neither takeaway requires the selector to be a new state of the art; both follow from treating example selection as an instrument and reporting what it moves. More broadly, the pattern here, an aggregate metric collapsing across a subgroup where the interesting behaviour lives, is not unique to ADR extraction, and the same stratify-by-expression lens should transfer to other extraction tasks whose targets range from canonical to colloquial.

8 Limitations

Our corpora are patient forums, drug reviews and biomedical case reports, not micro-blogs, and the task is span extraction only, without normalization to a medical dictionary. The cross-corpus comparison of Section 6 uses a single extraction model and selection strategy so that the corpora are what differ; we have not repeated it for all four models. PHEE’s implicit stratum is small (77 spans), so its interval is correspondingly wide. The register labeller’s human validation was carried out on the two

primary corpora; its labels on CADECv2 and PHEE were not independently human-validated, so we describe it as validated on the primary corpora and transferred to the others.

The stratification is LLM-generated, validated against human judgment on a 180-span sample (labelled by the authors, who were not blind to the hypothesis) rather than fully human-annotated across the test set. The boundary-sensitivity and cross-family checks (Sections 5.3 and 5.5) are our defence that neither borderline placement nor the labeller’s model family drives the result; a large representative human-labelled stratification would settle it more directly and is the natural next annotation study.

9 Conclusions

Four findings sit behind the single F1 score usually reported for this task, and they are clearer stated apart. First, and most important, accuracy within one model depends heavily on how a reaction is worded. Reactions put in everyday, figurative language are recovered far less often than clinically named ones, and aggregate F1 hides this completely. Second, a training-free, CPU-only selector that guarantees scarce reaction-bearing examples in the prompt recovers part of that colloquial deficit, precision-preserving on reaction-bearing posts, at a quantified false-alarm cost on reaction-free posts.

Third, the comparison with fine-tuning is narrower than the aggregate implies. A paired comparison on the same test items detects no difference between the selector and a supervised encoder in either register, although the selector uses no annotation or training beyond the labelled demonstration pool. That encoder’s remaining advantage in aggregate F1 therefore rests principally on suppressing false positives on reaction-free posts (about 7% against 26%), with a smaller contribution from higher precision on reaction-bearing posts. The encoder also still leaves a comparable implicit deficit, so the residual is not a prompting artifact; it persists across the paradigms we evaluated.

Fourth, the deficit is not peculiar to the corpora it was measured on. Under one fixed condition across four corpora, spanning patient forums, drug reviews and clinician-written case reports, how often reactions are phrased colloquially varies more than threefold while how much harder they are to extract does not measurably differ.

The reusable lesson is the diagnostic lens. Stratify an extraction benchmark by how its target is expressed, and a single F1 number resolves into a map of where a method helps, where it trades one error for another, and where the difficulty persists across every model, paradigm and corpus we evaluated. Such extractors are decision-support for pharmacovigilance triage, not autonomous safety systems. Because a missed or hallucinated reaction carries clinical cost, the operating point should be chosen with a human reviewer downstream, and the register-stratified view is precisely what lets that choice be made with the implicit-mention blind spot in view.

Acknowledgements. The fine-tuning experiment used compute on the Digital Research Alliance of Canada.

Abbreviations.	ADR	Adverse drug reaction
	ICL	In-context learning
	FT	Fine-tuning
	BM25	Best Matching 25 (lexical retrieval function)
	SBERT	Sentence-BERT (dense sentence embeddings)
	CI	Confidence interval
	PABAK	Prevalence-adjusted bias-adjusted kappa
	OR	Odds ratio

Statements and Declarations

- **Funding.** This research received no external funding.
- **Competing interests.** The authors declare no competing interests.
- **Ethics approval and consent to participate.** Not applicable. The study uses only previously published, publicly available corpora, used in accordance with their licences, and involved no new data collection from human subjects.
- **Consent for publication.** Not applicable.
- **Data availability.** No new data were generated. All four corpora are previously published and publicly available from their original authors: CADEC (Karimi et al. 2015) and CADECv2, released with the MultiADE benchmark (Dai et al. 2024), from CSIRO’s data portal; PsyTAR with its source publication (Zolnoori et al. 2019); and PHEE from its authors (Sun et al. 2022). Each was obtained from its official distribution and used under its own terms. The document identifiers removed from CADECv2 to de-duplicate it against CADEC are listed in the repository.
- **Materials availability.** Not applicable.
- **Code availability.** Code for the selector, scorer, stratified evaluation, leakage audit, boundary-sensitivity and confound analyses, together with the register-label consistency audit and its sensitivity test, the cross-corpus register-gap analysis, the false-positive classification, the explicit/implicit annotation codebook and the register labels for all four corpora, is available at <https://github.com/pyqmax/adr-social>.
- **Author contribution.** Conceptualization, methodology, validation, and formal analysis, T.T. and A.T.; software, data curation, investigation, and visualization, T.T.; original-draft writing, T.T. and A.T.; review and editing, supervision, and project administration, A.T. Both authors read and approved the final manuscript.

Appendix A Worked Prompt Example

Table A1 shows a complete prompt built by the minority-aware selector (ours_pos) for one PsyTAR test sentence, together with the model’s output. The prompt is the anchor benchmark’s expert-persona template (Elbiach et al. 2025), a fixed system instruction followed by the $k = 5$ selected demonstrations as alternating user (post) and assistant (comma-separated gold reactions, or “none”) turns, then the target post as the final user turn. The demonstrations are the actual examples the selector chose for this item; the positive guarantee has placed three reaction-bearing demonstrations in the prompt, including figurative ones (“feeling like I was going to fall over”, “felt

like I was not myself”), which is exactly the behaviour the guarantee is designed to produce on an implicitly worded target. The model recovers the two explicit reactions exactly and produces a truncated version of the figurative span, which strict scoring counts as a miss, a concrete instance of the implicit deficit the paper quantifies.

Table A1 A complete ours_pos prompt and the model output for one PsyTAR test item (GPT-4o-mini, 5-shot). “POS”/“neg” marks whether each selected demonstration is reaction-bearing.

Role	Content
System	As an expert in NLP with extensive experience in biomedical text analysis, carefully analyse the following text to identify mentions of Adverse Drug Reactions (ADRs). Extract only the ADR terms explicitly stated in the text, ensuring accuracy and completeness. List them clearly, separated by commas, without adding any extra words, explanations, or context. Focus solely on ADR terms to maximize precision and relevance.
User (demo 1, POS)	I just can’t go on indefinitely feeling like I was going to fall over or vomit all day.
Assistant	feeling like I was going to fall over, vomit all day
User (demo 2, POS)	I was shaky and could not sleep.
Assistant	shaky, could not sleep
User (demo 3, POS)	I slept a lot, had no appetite whatsoever, lost weight, felt like I was not myself at all.
Assistant	slept a lot, had no appetite, lost weight, felt like I was not myself
User (demo 4, neg)	I wasn’t able to sleep at night because I always felt like I was going crazy.
Assistant	none
User (demo 5, neg)	Constantly felt like I was about to faint, felt like my brain was being zapped with an electrical discharge.
Assistant	none
User (target)	shaky, weak, felt like I was going to faint if I stood at all.
Assistant (output)	shaky, weak, felt like I was going to faint

Gold reactions for the target: “shaky” (explicit), “weak” (explicit), “felt like I was going to faint if I stood at all” (implicit). Strict scoring credits the two explicit spans and counts the truncated figurative span as a miss.

Appendix B Per-Model Stratified Results

Table B2 gives the full item-level strict precision, recall, and F1 by register stratum for all four models, expanding the single implicit-F1 delta column of Table 5. Across every model, the minority guarantee (ours_pos) improves implicit F1 over BM25 (the paired significance is in Table 5), and where explicit extraction still has headroom (Llama, +6.0 explicit F1) it helps there too, while on the model with the least explicit

headroom (GPT-4o, +1.3) the gain concentrates on the implicit stratum. Negative-post false-alarm rates are reported for context. The guarantee’s false-alarm cost is largest on GPT-4o-mini (+5.7 points) and smaller on the other three, between -1.1 and $+3.3$.

Table B2 Per-model item-level strict P/R/F1 by register on PsyTAR (5-shot, validated LLM labels), and the negative-post false-alarm rate (FA). Explicit $n = 528$ spans / 159 items; implicit $n = 222$ spans / 182 items; 850 negative items. Paired significance for the implicit ours_pos-vs-bm25 gain is in Table 5.

Model	Selector	Explicit			Implicit			Neg
		P	R	F1	P	R	F1	FA%
GPT-4o-mini	random	68.1	69.5	68.8	69.5	55.1	61.5	18.2
	bm25	74.7	75.9	75.3	74.0	62.9	68.0	20.1
	ours_pos	77.0	79.7	78.4	73.8	67.9	70.7	25.8
GPT-4o	random	75.2	75.2	75.2	76.3	64.7	70.0	14.0
	bm25	78.1	76.8	77.5	80.2	68.1	73.6	12.1
	ours_pos	79.9	77.8	78.8	81.2	71.8	76.2	14.5
Llama-3.3-70B	random	64.5	66.6	65.5	58.4	53.8	56.0	30.2
	bm25	64.0	69.8	66.8	64.2	60.1	62.1	33.2
	ours_pos	71.2	74.6	72.8	66.6	61.7	64.1	32.1
Qwen-2.5-72B	random	72.0	74.6	73.3	68.4	63.1	65.6	15.8
	bm25	74.8	77.5	76.1	75.9	70.8	73.3	17.2
	ours_pos	77.8	83.6	80.6	74.6	76.3	75.5	20.5

Appendix C Consolidated Metrics

Table C3 collects, in one place, every quantity used in the main text for the primary model on PsyTAR, so that the span-level and item-level views can be read side by side without cross-referencing. Overall strict F1 is span-level (micro over posts); the per-stratum figures are span-recall and item-level F1 as defined in Section 3.1; the false-alarm rate is over the 850 reaction-free posts. The pattern is the one the main text describes. Retrieval lifts everything over random, and the minority guarantee adds recall on the implicit stratum at a false-alarm cost on negatives.

Appendix D Register Annotation Codebook

This is the codebook the two annotators followed (condensed; the full version and the annotation spreadsheets are in the code repository). The label describes *how the reaction is expressed*, not whether any model recovered it and not its severity or rarity. The examples below are generic illustrations chosen to define the two categories,

Table C3 Consolidated metrics, GPT-4o-mini on PsyTAR (5-shot, validated register labels, strict). Explicit stratum $n = 528$ spans; implicit $n = 222$; negatives 850 posts.

Selector	Overall	Span-recall		Item F1		Neg
	F1	expl.	impl.	expl.	impl.	FA%
random	55.31	72.3	34.2	68.8	61.5	18.2
bm25	59.74	78.2	44.6	75.3	68.0	20.1
ours_pos	59.80	81.8	51.4	78.4	70.7	25.8

Overall F1 with 95% CIs is in Table 4; per-model versions of the stratum columns are in Table B2; the fine-tuned band is in Table 7.

not selected by inspecting the test data. Common symptom names such as “nausea” inevitably also occur there; the figurative examples occur in neither test set, so annotators were not primed with the spans they would later label.

Explicit. The span names a recognized symptom using conventional clinical or standard lay vocabulary, recognizable as a bodily reaction with no inference beyond vocabulary. Examples: “nausea”, “headache”, “insomnia”, “dizziness”, “weight gain”; standard lay synonyms count (“throwing up”, “can’t sleep”).

Implicit. The span conveys a reaction through colloquial, figurative, vague, or experiential language that requires inference, metaphor, or world knowledge to recognize as a drug reaction. Examples: “feeling off”, “wiped out”, “like a zombie”, “not myself”, “couldn’t get out of bed”, “my brain is fog”.

Decision rule (in order). (1) If the span is a conventional symptom term or standard lay synonym, label EXPLICIT. (2) If recognizing it as a bodily reaction requires inference beyond vocabulary, label IMPLICIT. (3) For a mixed span, label by the head reaction word. Negated or hypothetical mentions are not reaction spans and are excluded; severity modifiers do not change the label.

Borderline cases for the sensitivity analysis of Section 5.3 are identified separately, as the spans on which the two validated labellers disagree, rather than from an annotator flag.

Appendix E Reproducibility Details

Data and splits: CADEC uses the benchmark test split (Elbiach et al. 2025) (188 posts); PsyTAR uses the repository split (train 3850 / test 1191 sentences). For Section 6, CADECv2 uses the official MultiADE .id split with every document also occurring in CADEC removed from both halves (161 train and 46 test documents dropped, leaving 2678 train / 663 test), matching on case-folded alphanumeric document text; PHEE uses the released `train.json/test.json` split (2898 train / 968 test). Gold spans in both added corpora are register-labelled by the same validated labeller and prompt as the primary corpora. Prompt: the anchor’s expert-persona template, shown in full in Appendix A; five demonstrations, temperature 0. Extraction models are queried through an OpenAI-compatible API (OpenRouter) as `openai/gpt-4o-mini`, `openai/gpt-4o`, `meta-llama/llama-3.3-70b-instruct`, `qwen/qwen-2.5-72b-instruct`; the register labeller is `openai/gpt-4.1-mini`, with

openai/gpt-4o as the second labeller for the boundary analysis of Section 5.3 and qwen/qwen-2.5-72b-instruct for the cross-family check of Section 5.5. Retrieval: BM25 is Okapi weighting (rank_bm25) over lowercased alphanumeric tokens; dense retrieval and clustering use all-MiniLM-L6-v2 sentence embeddings; k-means uses $\min(k, n)$ clusters, 10 restarts, seed 42, taking the example nearest each centroid. Selector: $n_{\text{pos}} = 3$ of $k = 5$, hard-pool $M = 50$ (Algorithm 1). Scoring: predictions are split on commas/semicolons/newlines, normalized (lowercased, whitespace-collapsed), and matched one-to-one greedily; strict requires exact match, relaxed accepts embedding cosine ≥ 0.8 ; empty and malformed outputs are scored as no prediction. Statistics: percentile bootstrap over test items, $B = 5000$; paired comparisons resample identical item indices for both conditions; comparisons against the fine-tuned encoders additionally resample the five training seeds; the pooled cross-model test resamples items shared across models. Fine-tuning: AutoModelForTokenClassification (PubMedBERT-base and Bio_ClinicalBERT) as a BIO tagger, 5 epochs, learning rate 3×10^{-5} , batch 16, maximum length 192, AdamW (defaults), seeds 42 to 46, BIO alignment against the extraction loader asserted before training. All code, prompts, labels, and the codebook are available (see Data Availability).

Appendix F Paired-Difference Intervals for Table 4

The aggregate comparisons of Table 4 are exploratory contrasts between selectors and are not Holm-adjusted. Their paired-difference intervals are given here (95%, $B = 5000$ paired bootstrap over test items).

Table F4 Paired-difference 95% intervals for the deltas of Table 4.

Corpus	Contrast	Δ F1	95% CI
CADEC	bm25 vs. random	+4.11	[+1.69, +6.46]
	ours_pos vs. random	+5.33	[+2.44, +8.15]
	sbert vs. random	+3.23	[+0.93, +5.60]
	cluster vs. random	−4.18	[−7.21, −1.39]
	ours_full vs. random	+2.51	[+0.16, +4.90]
	ours_pos vs. bm25	+1.23	[−0.54, +3.01]
	ours_full vs. bm25	−1.60	[−3.59, +0.34]
	cluster vs. bm25	−8.29	[−11.16, −5.53]
PsyTAR	bm25 vs. random	+4.43	[+1.56, +7.35]
	cluster vs. random	+4.61	[+2.68, +6.59]
	ours_pos vs. random	+4.49	[+1.48, +7.53]
	ours_full vs. random	+4.73	[+1.80, +7.75]
	ours_pos vs. bm25	+0.06	[−1.77, +1.95]
	ours_full vs. bm25	+0.30	[−1.44, +2.10]
	cluster vs. bm25	+0.18	[−2.66, +2.91]

References

- Alsentzer E, Murphy J, Boag W, et al (2019) Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp 72–78
- Baroian A (2025) Supervised fine-tuning or in-context learning? evaluating LLMs for clinical NER. arXiv preprint arXiv:251022285
- Breden A, Moore L (2020) Detecting adverse drug reactions from Twitter through domain-specific preprocessing and BERT ensembling. arXiv preprint arXiv:200506634
- Brown TB, Mann B, Ryder N, et al (2020) Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS)
- Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46(5):423–429. [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
- Chakma A, Surdeanu M, Blanco E (2026) Structured semantic information helps retrieve better examples for in-context learning applied to few-shot relation extraction. arXiv preprint arXiv:260120803
- Chandra M, Sriraman S, Verma G, et al (2025) Lived experience not found: LLMs struggle to align with experts on addressing adverse drug reactions from psychiatric medication use. In: 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp 11083–11113, <https://doi.org/10.18653/v1/2025.naacl-long.553>
- Chen H, Zhao Y, Chen Z, et al (2024) Retrieval-style in-context learning for few-shot hierarchical text classification. *Transactions of the Association for Computational Linguistics* 12:1214–1231. https://doi.org/10.1162/tacl_a.00697
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37–46. <https://doi.org/10.1177/001316446002000104>
- Dai X, Karimi S, Sarker A, et al (2024) MultiADE: A multi-domain benchmark for adverse drug event extraction. *Journal of Biomedical Informatics* 160:104744. <https://doi.org/10.1016/j.jbi.2024.104744>
- Dettmers T, Pagnoni A, Holtzman A, et al (2023) QLoRA: Efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems (NeurIPS)
- Dong Q, Li L, Dai D, et al (2024) A survey on in-context learning. arXiv preprint arXiv:230100234
- D’Oosterlinck K, Remy F, Deleu J, et al (2023) BioDEX: Large-scale biomedical adverse drug event extraction for real-world pharmacovigilance. In: Findings of

- the Association for Computational Linguistics: EMNLP 2023, pp 13425–13454, <https://doi.org/10.18653/v1/2023.findings-emnlp.896>
- Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman and Hall/CRC
- Elbiach O, Grissette H, Nfaoui EH (2025) Benchmarking large language models for adverse drug reaction extraction in social media and clinical texts. Results in Engineering 28:107362. <https://doi.org/10.1016/j.rineng.2025.107362>
- Golder S, Norman G, Loke YK (2015) Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. British Journal of Clinical Pharmacology 80(4):878–888. <https://doi.org/10.1111/bcp.12746>
- Grattafiori A, et al (2024) The Llama 3 herd of models. arXiv preprint arXiv:240721783
- Gu Y, Tinn R, Cheng H, et al (2021) Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare 3(1):1–23. <https://doi.org/10.1145/3458754>
- Gu Y, Zhang S, Usuyama N, et al (2023) Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. arXiv preprint arXiv:230706439
- Gwet KL (2008) Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology 61(1):29–48. <https://doi.org/10.1348/000711006X126600>
- Harris J, Laurence T, Loman L, et al (2024) Evaluating large language models for public health classification and extraction tasks. arXiv preprint arXiv:240514766
- Hu EJ, Shen Y, Wallis P, et al (2022) LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR)
- Karimi S, Metke-Jimenez A, Kemp M, et al (2015) CADEC: A corpus of adverse drug event annotations. Journal of Biomedical Informatics 55:73–81. <https://doi.org/10.1016/j.jbi.2015.03.010>
- Li H, Lin H (2026) Adverse drug reaction detection on social media based on large language models. Information 17(4):352. <https://doi.org/10.3390/info17040352>
- Li H, Zhang Y, Zhang Y, et al (2024) SRCB at #SMM4H 2024: Making full use of LLM-based data augmentation in adverse drug event extraction and normalization. In: 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks (SMM4H), pp 32–37, <https://doi.org/10.18653/v1/2024.smm4h-1.8>

- Liu J, Shen D, Zhang Y, et al (2022) What makes good in-context examples for GPT-3? In: Deep Learning Inside Out (DeeLIO 2022), pp 100–114, <https://doi.org/10.18653/v1/2022.deelio-1.10>
- Lu Y, Bartolo M, Moore A, et al (2022) Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp 8086–8098, <https://doi.org/10.18653/v1/2022.acl-long.556>
- Magge A, Klein A, Miranda-Escalada A, et al (2021) Overview of the sixth social media mining for health applications (#SMM4H) shared tasks. In: Sixth Social Media Mining for Health Applications Workshop and Shared Tasks (SMM4H), pp 21–32, <https://doi.org/10.18653/v1/2021.smm4h-1.4>
- Min S, Lyu X, Holtzman A, et al (2022) Rethinking the role of demonstrations: What makes in-context learning work? In: 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 11048–11064, <https://doi.org/10.18653/v1/2022.emnlp-main.759>
- Nikfarjam A, Sarker A, O’Connor K, et al (2015) Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22(3):671–681. <https://doi.org/10.1093/jamia/ocu041>
- OpenAI (2024) GPT-4o system card. Available online: <https://openai.com> (accessed on 1 July 2026)
- Peng K, Ding L, Yuan Y, et al (2024) Revisiting demonstration selection strategies in in-context learning. In: 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers, pp 9090–9101, <https://doi.org/10.18653/v1/2024.acl-long.492>
- Qwen Team (2024) Qwen2.5 technical report. arXiv preprint arXiv:241215115
- Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), <https://doi.org/10.18653/v1/D19-1410>
- Robertson S, Zaragoza H (2009) The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 3(4):333–389. <https://doi.org/10.1561/15000000019>
- Rocchetti M (2026) Before the algorithm: An exemplar case of the necessity of statistical testing for epidemiological consistency in public health data. *AIMS Public Health* 13(1):121–134. <https://doi.org/10.3934/publichealth.2026008>

- Rubin O, Herzig J, Berant J (2022) Learning to retrieve prompts for in-context learning. In: 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp 2655–2671, <https://doi.org/10.18653/v1/2022.naacl-main.191>
- Sarker A, Ginn R, Nikfarjam A, et al (2015) Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics* 54:202–212. <https://doi.org/10.1016/j.jbi.2015.02.004>
- Sun Z, Li J, Pergola G, et al (2022) PHEE: A dataset for pharmacovigilance event extraction from text. In: 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 5571–5587, <https://doi.org/10.18653/v1/2022.emnlp-main.376>
- Sun Z, Pergola G, Wallace BC, et al (2024) Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In: 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Volume 2: Short Papers, pp 344–357, <https://doi.org/10.18653/v1/2024.eacl-short.30>
- Walsh J, Mamidanna S, Nye B, et al (2025) Fine-tuning for better few-shot prompting: An empirical comparison for short answer grading. In: 2nd Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC 2025)
- Wei J, Wang X, Schuurmans D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. In: *Advances in Neural Information Processing Systems* (NeurIPS)
- Yazdani A, Rouhizadeh H, Vicente Alvarez D, et al (2023) DS4DH at #SMM4H 2023: Zero-shot adverse drug events normalization using sentence transformers and reciprocal-rank fusion. In: 8th Social Media Mining for Health Research and Applications Workshop and Shared Tasks (SMM4H)
- Yu S, Bao R, Bhatia P, et al (2025) Dynamic uncertainty ranking: Enhancing retrieval-augmented in-context learning for long-tail knowledge in LLMs. In: 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pp 8985–8997, <https://doi.org/10.18653/v1/2025.naacl-long.453>
- Zeng QT, Tse T (2006) Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association* 13(1):24–29. <https://doi.org/10.1197/jamia.M1761>
- Zheng Y, Gong J, Ren S, et al (2024) LHS712.ADENotGood at #SMM4H 2024 task 1: Deep-LLMADEminer: A deep learning and LLM pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. In: 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks (SMM4H), pp 130–132, <https://doi.org/10.18653/v1/2024.smm4h-1.30>

Zhong Q, Ding L, Liu J, et al (2023) Can ChatGPT understand too? a comparative study on ChatGPT and fine-tuned BERT. arXiv preprint arXiv:230210198

Zolnoori M, Fung KW, Patrick TB, et al (2019) A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. Journal of Biomedical Informatics 90:103091. <https://doi.org/10.1016/j.jbi.2018.12.005>