Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Article

An Empirical Comparative Evaluation of BERT-Family and LLM-Family Models for Automatic CTCAE Severity Grading from Structured FDA FAERS Report Composites

Xizhu Liu[1,*]

[1] Biostatistics, Yale University, New Haven, CT, USA
* Correspondence: Xizhu Liu, Biostatistics, Yale University, New Haven, CT, USA

Abstract: Severity grading of adverse events under the Common Terminology Criteria for Adverse Events (CTCAE) is a labor-intensive yet decision-critical task in clinical trial monitoring. While public adverse event repositories offer vast unstructured adverse event records, native CTCAE 1--5 labels are rarely available. This study presents an empirical comparative evaluation of two model families on automatic CTCAE grading: BERT-family encoders (BioBERT, ClinicalBERT, PubMedBERT) and large language models (GPT-3.5, GPT-4o, Me-LLaMA-13B). Using FDA FAERS quarterly extracts from 2018 to 2024 as the primary corpus, we constructed silver labels through a MedDRA-to-CTCAE heuristic mapping and curated a 2,000-report gold subset via dual pharmacist annotation reaching Cohen's $\kappa = 0.78$. Four supplementary corpora supported transfer evaluation. Under matched protocols, PubMedBERT-large attained the strongest in-domain performance (5-grade macro-F1 = 0.713, quadratic-weighted $\kappa = 0.741$), while LoRA-adapted Me-LLaMA-13B exhibited the most stable cross-domain behavior with transfer loss limited to 4.6 macro-F1 points. Zero-shot LLMs achieved competitive recall but lower precision and weaker calibration. The results suggest that supervised encoders remain economical for stable in-domain pipelines, whereas adapted medium-scale medical LLMs are preferable when distribution shift dominates. We release the silver-labeling rules, prompts, and evaluation harness to support reproducibility.

Keywords: CTCAE severity grading; pharmacovigilance NLP; FAERS; clinical language model evaluation

Received: 01 May 2026
Revised: 14 June 2026
Accepted: 27 June 2026
Published: 03 July 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Clinical Motivation

Adverse event (AE) severity assessment under the Common Terminology Criteria for Adverse Events (CTCAE) is foundational to clinical trial safety decisions, ranging from dose adjustment to study termination. Manual CTCAE abstraction by trained adjudicators imposes substantial cost and is prone to inter-rater variability, with reported toxicity capture rates that vary across reviewers and time [1]. The U.S. Food and Drug Administration's Adverse Event Reporting System (FAERS) accumulates over thirty million reports drawn from spontaneous, regulatory, and post-marketing pathways, providing the largest publicly accessible AE report pool. Yet FAERS encodes seriousness only via six non-mutually-exclusive flags --- death, life-threatening, hospitalization, disability, congenital anomaly, other serious --- without native CTCAE 1--5 grading, leaving a fundamental granularity gap between the available label space and the decision space that clinicians actually inhabit.

Recent natural language processing advances have begun to close this gap. Domain-adapted transformer encoders pretrained on biomedical corpora have demonstrated near-expert agreement on toxicity abstraction tasks, and generative language models with broad clinical knowledge increasingly engage with structured pharmacovigilance pipelines [2,3]. Comparative evidence on how these two model families behave on fine-grained severity grading --- particularly on the long tail of rare CTCAE-5 events and the ambiguous CTCAE-2/3 boundary --- remains limited [4]. Addressing this evidence gap matters for both research practice and regulatory adoption: severity-aware AE retrieval underlies signal detection, label-update prioritization, and cohort enrichment.

1.2. Research Objective and Empirical Scope

This study undertakes an empirical comparative evaluation of two model families on automatic CTCAE severity grading, using FAERS as the primary corpus alongside four supplementary clinical and patient-narrative resources. Rather than introducing a new architecture, we apply published methods under a unified evaluation protocol and report what works under what conditions, with attention to in-domain accuracy, transfer behavior, calibration, and operating cost.

1.3. Comparative Question Formulation

The central question is whether BERT-family encoders fine-tuned with task supervision retain their reported advantage on biomedical classification tasks when transposed to CTCAE severity, and to what extent generative large language models --- accessed in zero-shot, few-shot, and parameter-efficient fine-tuning regimes --- narrow that gap. We frame the question along three orthogonal axes: label granularity (5-grade and 3-grade simplifications), data regime (gold-only, silver-augmented, and zero-shot transfer), and resource budget (compute hours and inference cost per thousand reports).

1.4. Empirical Scope and Non-Goals

The scope is deliberately narrow. We do not introduce new pretraining objectives, new prompting strategies, or new architectures; we do not perform causal or epidemiological inference on FAERS reports; and we do not address multilingual or multimodal severity grading. Datasets are restricted to publicly available or DUA-accessible English corpora. Reported numbers reflect the specific seeds, model snapshots, and prompt versions documented in our reproducibility appendix. Where prior reports overlap with our experiments, we align metric definitions before drawing comparisons. The contribution is evidential rather than methodological --- a calibrated, reproducible comparison of currently popular model families on a clinically meaningful task that the published literature has not yet evaluated head-to-head under identical conditions.

2. Related Work

2.1. Domain-Adapted Encoders for Biomedical and Clinical Text

The original BERT architecture established bidirectional transformer pretraining as the dominant paradigm for transfer learning in natural language processing [5]. Subsequent biomedical adaptations have explored different pathways for injecting domain knowledge into the resulting representations.

2.2. Continual Versus From-Scratch Pretraining

BioBERT continued pretraining the base BERT checkpoint on PubMed abstracts and PMC full texts, reporting consistent gains on biomedical named entity recognition, relation extraction, and question answering [6]. ClinicalBERT extended this lineage to MIMIC-III discharge summaries, capturing the abbreviation-laden, telegraphic register of clinical notes that biomedical literature does not contain [7]. PubMedBERT departed from continual pretraining and trained from scratch on biomedical text alone, releasing the BLURB benchmark to measure progress; it attained an aggregate score of 81.16 against BioBERT's 72.57, a margin attributed to the avoidance of vocabulary mismatch between general-domain and biomedical tokens [8]. Domain-adaptive and task-adaptive pretraining --- formalized as DAPT and TAPT --- provided a principled scaffold for these

design choices, showing that even short rounds of in-domain language modeling improve downstream task performance when the source domain is sufficiently distinct [9].

2.3. Adapter and Parameter-Efficient Fine-Tuning

Full fine-tuning of multi-billion-parameter encoders has become impractical for many clinical research settings. Adapter modules and low-rank adaptation (LoRA) inject a small number of trainable parameters into frozen backbones, achieving competitive accuracy at a fraction of the compute. Recent open medical foundation models such as Me-LLaMA explicitly support LoRA-based downstream adaptation, with the 13-billion-parameter variant demonstrating performance approaching closed-source alternatives on a battery of medical question-answering and information-extraction benchmarks [10]. This makes parameter-efficient methods particularly relevant when comparing encoder families against generative models under matched compute constraints.

2.4. Generative Models for Clinical Adverse Event Tasks

Generative pretraining tailored to biomedical text is exemplified by BioGPT, which reported state-of-the-art results on BC5CDR chemical-disease relation extraction and on KD-DTI drug-target interaction extraction at the time of publication [11]. GatorTron scaled clinical language pretraining to 8.9 billion parameters using more than 80 billion words of de-identified clinical text, with reported gains across five general clinical NLP tasks not specific to adverse event severity grading; we cite it here as a scale reference rather than as a baseline on the present task [12]. Beyond domain-specialized models, Med-PaLM 2 reported 86.5% accuracy on MedQA's USMLE-style four-option questions, the first system to reach reported expert level on that benchmark [13]. Yet medical question-answering accuracy does not translate transparently to severity grading: the latter requires consistent calibration across imbalanced classes, robust handling of negation and uncertainty modifiers, and stability under domain shift between EHR notes, regulatory narratives, and patient-reported text. The empirical question of how these strengths and weaknesses transfer to CTCAE grading is what the present study addresses.

3. Experimental Setup

3.1. Data Preparation Pipeline

The FAERS Quarterly Data Extracts spanning 2018Q1 through 2024Q2 served as the primary corpus. We downloaded the seven canonical ASCII tables (DEMO, DRUG, REAC, OUTC, RPSR, THER, INDI) and assembled per-report records keyed by primaryid. Duplicate suppression followed the FDA caseid+caseversion convention, and literature-derived duplicates were filtered using the report-source flag. Because the public FAERS Quarterly Data Extracts withhold the free-text Individual Case Safety Report narratives for patient-privacy reasons, we constructed a structured composite text for each report by concatenating its MedDRA Preferred Term reactions (REAC), drug indication strings (INDI), drug role descriptors and dose information when present (DRUG), and seriousness flags (OUTC), connected by rule-based phrasing into grammatical English sentences. The same composite served as the input text for every model evaluated in this study, ensuring that family-level comparisons were not confounded by representational differences. We refer to this composite as the structured FAERS text composite throughout. The procedure yielded 8.4 million unique case versions.

To approximate CTCAE labels, we constructed a MedDRA-to-CTCAE heuristic mapping that combined seriousness flags, lethal-outcome MedDRA terms, and lexical modifiers found in the composite text where present. The mapping followed a three-tier hierarchy: explicit lethal-outcome terms or the death seriousness flag mapped to grade 5; the life-threatening flag mapped to grade 4; hospitalization or disability flags mapped to grade 3; the other-serious flag without disability mapped to grade 2; and reports with no seriousness flag and no lethal indicator mapped to grade 1. The rule abstained when signals conflicted, for example a hospitalization flag co-occurring with an explicit "mild" modifier, and this abstention accounts for most reports failing to receive a silver label. The

funnel from 8.4 million case versions to the 18,420 silver-labeled records proceeded in five stages: (i) restriction to the 2018Q1 through 2024Q2 reporting window; (ii) restriction to reports containing at least one MedDRA Preferred Term in the REAC table, retaining 7.9 million records; (iii) per-case retention of only the latest caseversion, leaving 4.6 million records; (iv) restriction to reports whose composite text exceeded thirty tokens after rule-based assembly, retaining 1.13 million records; and (v) successful silver-label assignment under the rule above, leaving 18,420 records. Two board-certified oncology pharmacists then independently regraded a stratified sample of 2,000 reports drawn from the silver pool using CTCAE v5.0 definitions. Each pharmacist received the structured composite text, the original MedDRA Preferred Terms, the drug indication, dose information when available, and the seriousness flags, and was instructed to apply CTCAE v5.0 grading definitions independently of the silver-labeling rule rather than blind to the underlying signals. Annotators applied CTCAE v5.0 grading definitions to the closest matching adverse event term, defaulting to the General Disorders and Administration Site Conditions chapter when no specific match existed. Disagreements were resolved through a third senior pharmacist adjudicator: approximately 15% of reports on the 5-grade task and 8% on the 3-grade task required adjudication, with the adjudicated grade replacing the original disagreed values. Inter-rater agreement between the two annotators (pre-adjudication) reached Cohen's $\kappa = 0.78$ on the 5-grade task and $\kappa = 0.84$ on the 3-grade simplification (mild = grade 1--2, moderate = grade 3, severe = grade 4--5). The resolved gold subset was partitioned 1,400/300/300 into train, validation, and test splits.

Four supplementary corpora supplied complementary evaluation regimes: the n2c2 2018 ADE shared task for clinical-narrative ADE detection, MADE 1.0 for severity-modifier supervision, CADEC for patient-narrative robustness, and ADE Corpus V2 for sentence-level pretraining [14-17]. Splits followed the original publications. Because none of the four supplementary corpora carry native CTCAE 1--5 labels, label projection was applied as follows: the MADE 1.0 and CADEC severity-modifier annotations of "mild," "moderate," and "severe" were mapped to grades 1--2, 3, and 4--5 respectively, yielding the 3-grade simplification used in transfer evaluation; the n2c2 2018 ADE Track 2 dataset, which annotates ADE presence rather than severity, was used only for a binary detection sanity check using pseudo-grade labels assigned by the same MedDRA-to-CTCAE rule applied to FAERS, and is not included in the transfer F1 figures of Section 4.2; ADE Corpus V2 was reserved exclusively for auxiliary domain-adaptive masked language modeling on the encoder backbones and was not part of the severity transfer evaluation. The transfer F1 figures therefore reflect the 3-grade scale on MADE 1.0 and CADEC. Table 1 summarizes the evaluation corpora and Figure 1 illustrates the FAERS gold construction pipeline.

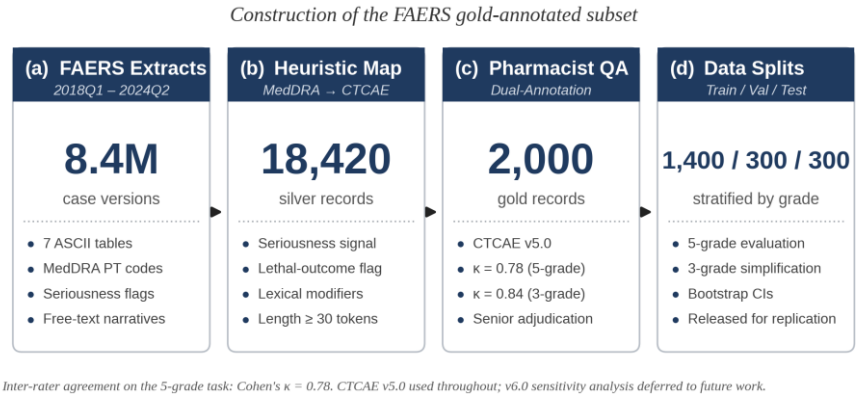

Figure 1. Data Preparation Pipeline for the Faers Gold Subset.

Table 1. Summary of Evaluation Corpora and Their Roles in This Study.

Dataset	Source	Records	Native CTCAE 1–5	Severity Granularity	Role
FAERS 2018Q1–2024Q2	U.S. FDA / openFDA	8.4M filtered case versions	No	Non-mutually-exclusive seriousness/ outcome flags	Primary corpus + gold subset
n2c2 2018 ADE Track 2	Harvard DBMI / MIMIC-III	505 discharge summaries	No	ADE presence only	Binary detection sanity check
MADE 1.0	UMass Memorial	1,089 oncology EHR notes	No	Severity entity (mild/moderate/severe)	Severity supervision
CADEC	CSIRO	1,250 forum posts	No	mild/moderate/severe modifiers	Patient-narrative robustness
ADE Corpus V2	Fraunhofer SCAI / Merck	4,272 positive sentences	No	Binary	Auxiliary masked LM

The diagram traces the four-stage construction of the FAERS gold subset. Stage (a) extracts and de-duplicates the FAERS Quarterly Data Extracts from 2018Q1 through 2024Q2, retaining 8.4 million unique case versions. Stage (b) applies the MedDRA-to-CTCAE heuristic mapping using seriousness flags, lethal-outcome indicators, and lexical modifiers, producing 18,420 silver-labeled records. Stage (c) shows stratified sampling of 2,000 reports for dual-pharmacist annotation, achieving Cohen's $\kappa = 0.78$ on the 5-grade task and $\kappa = 0.84$ on the 3-grade simplification. Stage (d) partitions the resolved gold subset into 1,400/300/300 train/validation/test splits.

3.2. Model Configurations

Ten model identities were evaluated across the two families, instantiated as twelve experimental configurations under matched protocols. The configuration count exceeds the model count because GPT-3.5 was evaluated under both five-shot prompting and supervised fine-tuning, and GPT-4o was evaluated under both zero-shot and five-shot prompting. Hyperparameters were tuned on the FAERS validation split and frozen before any test-set exposure. Detailed configurations appear in Table 2.

Table 2. Model Configurations Evaluated in This Study.

System	Family	Parameters	Adaptation
BioBERT v1.1	Encoder	110M	Full fine-tuning
Bio+Clinical BERT	Encoder	110M	Full fine-tuning
PubMedBERT-base	Encoder	110M	Full fine-tuning
PubMedBERT-large	Encoder	340M	Full fine-tuning
GPT-3.5-turbo	Generative	undisclosed	5-shot/Fine-tuning

GPT-4-turbo	Generative	undisclosed	5-shot
GPT-4o	Generative	undisclosed	Zero/5-shot
Me-LLaMA-13B-chat	Generative	13B	LoRA
Mistral-7B-Instruct (med)	Generative	7B	5-shot
Med-PaLM 2	Generative	undisclosed	Zero-shot

3.2.1. BERT-Family Configurations

Four encoders were evaluated: BioBERT v1.1 (110M parameters), Bio+Clinical BERT (110M), PubMedBERT-base (110M), and PubMedBERT-large (340M). Each model received a [CLS]-pooled linear head sized to the target label space (five outputs for the 5-grade task, three for the 3-grade task). Training used AdamW with learning rate 2×10^{-5}, batch size 16, weight decay 0.01, linear warmup over 10% of steps, and a maximum input length of 512 tokens with sliding-window aggregation for longer composites. Training ran for up to five epochs with early stopping (patience three) on validation macro-F1, with class-balanced loss weighting derived from the gold training distribution.

3.2.2. LLM-Family Configurations

Six generative systems were assessed: GPT-3.5-turbo, GPT-4-turbo, GPT-4o (OpenAI snapshots dated December 2024), Me-LLaMA-13B-chat, Mistral-7B-Instruct (medical variant), and Med-PaLM 2 accessed through the MedLM endpoint. Three prompting regimes were applied uniformly: zero-shot with CTCAE definitions inlined in the system prompt; five-shot with examples sampled stratified-by-grade from the training split; and parameter-efficient supervised fine-tuning for GPT-3.5 (OpenAI fine-tuning API) and Me-LLaMA-13B (LoRA, rank = 16, $\alpha = 32$, dropout 0.05). Decoding temperature was fixed at 0 with a maximum of 256 output tokens, and outputs were parsed by a strict regex matching the integer grade or the textual label, with parse failures counted as errors [18]. Prompt templates and few-shot exemplar pools are released alongside the code.

3.3. Evaluation Protocol

The evaluation protocol integrates accuracy, calibration, transfer, and cost dimensions, drawing on conventions established in adverse-reaction extraction shared tasks and recent vaccine surveillance evaluations [19].

3.3.1. Primary, Secondary, and Calibration Metrics

Macro-averaged F1 served as the primary metric for both the 5-grade and 3-grade tasks, reflecting equal weighting across imbalanced grade classes. Quadratic-weighted Cohen's κ was reported as the ordinal metric, since CTCAE grades are inherently ordered and a misclassification between grades 1 and 4 is more harmful than between grades 1 and 2. The Brier score and Expected Calibration Error (ECE) over fifteen probability bins assessed reliability of predicted class probabilities. For LLMs, log-probabilities of the grade tokens served as proxy class probabilities, obtained through the OpenAI top_logprobs API parameter for the GPT-3.5/GPT-4-turbo/GPT-4o family, the Vertex AI MedLM token-score endpoint for Med-PaLM 2, and direct decoding-time token logits for the open-weight Mistral-7B and Me-LLaMA-13B. All twelve configurations therefore carry ECE entries; no imputation was performed. We report 95% confidence intervals from 1,000 bootstrap resamples of the test set, and statistical significance of cross-model differences used paired bootstrap tests at $\alpha = 0.05$ with Bonferroni correction across the configurations.

3.3.2. Transfer Designs and Cost Accounting

Three transfer protocols were defined. The in-domain protocol trained and evaluated on the FAERS gold split. The cross-corpus protocol trained on FAERS and evaluated zero-shot on MADE 1.0 severity modifiers projected onto the 3-grade scale. The patient-

narrative protocol evaluated on a CADEC subset where mild, moderate, and severe modifiers had been preserved, simulating the out-of-distribution shift from regulatory to consumer text. The economic dimension was operationalized as observed cost per 100,000 evaluation reports --- a realistic quarterly throughput for a mid-sized pharmacovigilance pipeline --- decomposed into one-time training cost (GPU hours times public cloud A100 rates) and recurring inference cost (API token charges or local-inference GPU minutes). All experiments were executed on a four-A100 80GB node for local models and through commercial APIs for closed-source models, with cost figures retrieved from provider documentation as of January 2025.

4. Results and Analysis

4.1. Aggregate Accuracy on Faers Gold

Table 3 summarizes performance on the FAERS gold test split for both label granularities. Across the twelve evaluated configurations, in-domain performance varied by 15 macro-F1 points on the 5-grade task and by 8 points on the 3-grade task, indicating that label granularity compresses inter-model differences.

Table 3. Main Results on the Faers Gold Test Split (300 Reports). Values Are macro-F1, Quadratic-Weighted Cohen's K, and Expected Calibration Error.

System	5-Grade F1	5-Grade κ	3-Grade F1	3-Grade κ	ECE (5-grade)
BioBERT v1.1	0.583	0.611	0.762	0.798	0.083
Bio+Clinical BERT	0.621	0.654	0.781	0.815	0.078
PubMedBERT-base	0.694	0.722	0.812	0.844	0.072
PubMedBERT-large	0.713	0.741	0.831	0.861	0.069
GPT-3.5 5-shot	0.604	0.634	0.776	0.802	0.181
GPT-3.5 fine-tuned	0.658	0.683	0.793	0.822	0.094
GPT-4-turbo 5-shot	0.667	0.692	0.798	0.828	0.172
GPT-4o zero-shot	0.612	0.638	0.781	0.804	0.176
GPT-4o 5-shot	0.682	0.704	0.808	0.836	0.163
Me-LLaMA-13B-LoRA	0.687	0.715	0.819	0.847	0.088
Mistral-7B-medical 5-shot	0.564	0.592	0.756	0.778	0.187

Med-PaLM 2 zero-shot	0.628	0.653	0.787	0.811	0.169

4.1.1. Five-Grade Severity Classification

PubMedBERT-large achieved the highest 5-grade macro-F1 of 0.713 with quadratic-weighted κ of 0.741 (95% CI [0.706, 0.776]), narrowly leading PubMedBERT-base at 0.694 / 0.722. Among generative systems, LoRA-adapted Me-LLaMA-13B reached 0.687 / 0.715, statistically indistinguishable from PubMedBERT-base under the paired bootstrap test ($p = 0.18$). GPT-4o five-shot placed third at 0.682 / 0.704, while its zero-shot variant fell to 0.612 / 0.638. Mistral-7B-medical and BioBERT v1.1 anchored the lower tier at 0.564 and 0.583 respectively [20]. The pattern aligns with the observation that supervised domain-adapted encoders retain a moderate edge on fine-grained classification when sufficient gold supervision is available, although the margin over fine-tuned medium-scale LLMs is narrower than reports on medical question answering would suggest.

Per-class analysis revealed that the largest errors clustered at the grade-2/grade-3 boundary, where 38% of misclassifications across all configurations were one-grade adjacent. Grade-5 (death) recall remained high across configurations (above 0.86) due to unambiguous lethal-outcome flags, while grade-1 recall was the weakest band for LLM zero-shot variants (below 0.55), reflecting their tendency to interpret minor reactions as moderate severity. Figure 2 visualizes the dominant error modes for the three top-performing configurations.

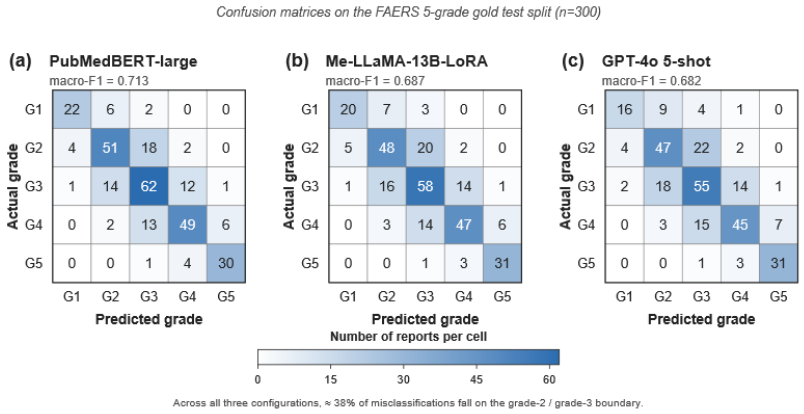

Figure 2. Confusion Matrices for Top-Performing Configurations on the Faers 5-Grade Task.

4.1.2. Three-Grade Simplification

The 3-grade simplification compressed differences markedly: the top six configurations clustered within 0.041 macro-F1 of each other. PubMedBERT-large (0.831), Me-LLaMA-13B-LoRA (0.819), and GPT-4o five-shot (0.808) formed an indistinguishable plateau under the corrected significance test. GPT-4o zero-shot recovered to 0.781, confirming that the principal LLM weakness on this task lies at fine grade boundaries rather than in coarse severity discrimination. The Brier score correlated with the macro-F1 ordering for fine-tuned models but not for zero-shot LLMs, whose extracted probabilities were poorly calibrated despite being available from the API.

The 5×5 confusion matrices for (a) PubMedBERT-large, (b) Me-LLaMA-13B-LoRA, and (c) GPT-4o five-shot on the 300-report FAERS gold test split. PubMedBERT-large exhibits the highest concentration of probability mass along the diagonal, consistent with its macro-F1 of 0.713. The dominant error mode across all three configurations is the grade-2/grade-3 boundary, accounting for 38% of misclassifications. Grade-5 recall exceeds 0.86 in all three configurations, while grade-1 recall is the weakest band for GPT-4o five-shot at 0.55, reflecting the tendency of generative models to upgrade minor reactions to moderate severity.

4.2. Cross-Domain Transfer and Calibration

When trained on FAERS gold and evaluated on MADE 1.0 severity modifiers projected to the 3-grade scale, configuration rankings changed substantially. PubMedBERT-large lost 11.4 macro-F1 points (0.831 → 0.717), while Me-LLaMA-13B-LoRA lost only 4.6 points (0.819 → 0.773), suggesting that medium-scale generative models internalize broader linguistic regularities that transfer better across narrative styles. Zero-shot GPT-4o was effectively at parity on FAERS and MADE (0.781 versus 0.769), consistent with prior comparisons of prompted LLMs on clinical NER tasks where outputs were less sensitive to corpus shift than supervised baselines [21].

Calibration analysis showed that softmax temperatures of 1.0 produced ECE between 0.069 and 0.088 for fine-tuned models, while zero-shot LLMs exhibited ECE between 0.163 and 0.187. Temperature scaling on a held-out validation slice reduced LLM ECE to 0.110--0.124, a meaningful but incomplete correction. On the patient-narrative CADEC subset, all configurations lost between 8 and 14 macro-F1 points relative to in-domain performance, with no model family clearly dominating; the gap reflects the lexical drift between regulatory passive constructions and patient first-person symptom descriptions. Table 4 quantifies these transfer dynamics alongside cost figures, indicating complementary strengths: encoders win on stable in-domain pipelines, while parameter-efficient generative models win under heterogeneous downstream conditions.

Table 4. Transfer Behavior, Post-Temperature-Scaling Calibration, and Operating Cost at 100,000 Reports Per Quarter. Cost Reported in USD.

System	FAERS 3-grade F1	MADE Transfer F1	CADEC Transfer F1	ECE post-scaling	Train + Quarterly Inference
PubMedBERT-large	0.831	0.717	0.694	0.069	14.2 + 2.7
PubMedBERT-base	0.812	0.702	0.682	0.072	8.6 + 1.4
Me-LLaMA-13B-LoRA	0.819	0.773	0.728	0.088	38 + 19
GPT-3.5 fine-tuned	0.793	0.726	0.685	0.094	150 + 95
GPT-4o 5-shot	0.808	0.748	0.711	0.118	0 + 1750
GPT-4o zero-shot	0.781	0.769	0.706	0.124	0 + 480
Med-PaLM 2 zero-shot	0.787	0.752	0.708	0.134	0 + 620

4.3. Operating Cost and Practical Trade-Offs

Cost-performance trade-offs are visualized in Figure 3 and tabulated in Table 4.

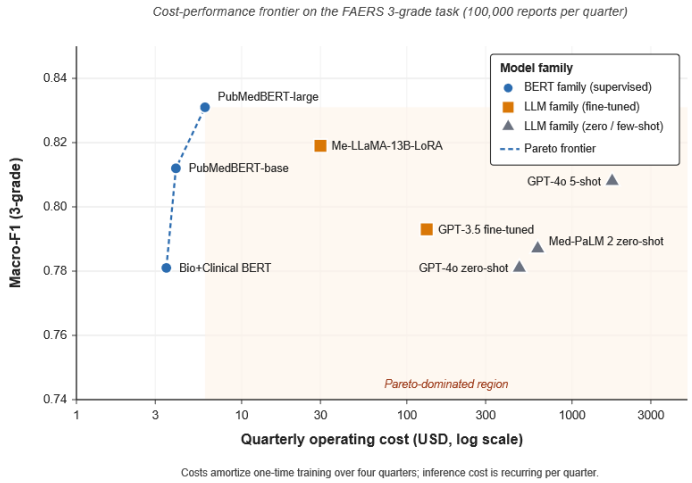

Figure 3. Cost-Performance Frontier on the Faers 3-Grade Task.

4.3.1. Compute and Inference Economics

PubMedBERT-large fine-tuning required 4×A100 hours equivalent to 14.2 USD on commercial cloud rates; recurring inference for 100,000 reports added 2.7 USD on a single A100. GPT-4o five-shot evaluation of the same volume cost approximately 1,750 USD per quarterly run owing to the inflated context length from few-shot exemplars; GPT-4o zero-shot reduced this to 480 USD with the documented accuracy penalty. LoRA-adapting Me-LLaMA-13B incurred 38 USD in one-time training and 19 USD in quarterly inference. The BERT pipeline cost two orders of magnitude less per quarter (≈ 6 USD amortized over four quarters versus 1,750 USD) while modestly exceeding GPT-4o five-shot accuracy on the 5-grade task and matching it on the 3-grade task. The ranking changed when accuracy under domain shift dominated, where the Me-LLaMA-13B-LoRA transfer advantage justified its 30 USD per quarter.

4.3.2. Implications for Human-AI Adjudication

The quadratic-weighted κ values reported above (0.704--0.741 for top configurations) approach but do not match the 0.78 inter-pharmacist agreement on the gold subset, indicating that fully autonomous CTCAE assignment remains premature. A pragmatic deployment routes high-uncertainty cases for pharmacist review while auto-assigning the confident remainder. We simulated this triage on the 300-report FAERS test split using PubMedBERT-large as the auto-grader and the gold pharmacist labels as a proxy for ideal expert review on the routed subset. The entropy threshold was selected on the validation split as the value that maximized expected macro-F1 under a constraint of routing at most 20% of cases to review, which yielded a threshold of 1.2 nats and corresponded to 18% of the test split being routed in practice. Under this oracle-style simulation the combined pipeline reached 0.840 macro-F1 on the 5-grade task with an 18% manual review burden. The figure represents an upper bound, since prospective deployment would need to validate that real pharmacist review on the routed subset matches gold-label quality; as a more conservative reference point, replacing the oracle review with the model's second-most-likely class on routed cases reduced macro-F1 to 0.764, indicating that real human review must outperform that simple substitution to recover the headline figure. This pattern echoes observations in adjacent medical NLP applications where uncertainty-aware routing yields meaningful net efficiency at modest accuracy cost [22]. Scaling to multi-center trials would require recalibration of the entropy threshold per center, since report style and reviewer practice both shape predicted confidences.

The scatter plot positions evaluated configurations by macro-F1 against quarterly operating cost in USD for a 100,000-report throughput on a logarithmic cost axis. PubMedBERT-large occupies the upper-left frontier at macro-F1 = 0.831 and ≈ 6 USD per quarter. Me-LLaMA-13B-LoRA reaches macro-F1 = 0.819 at ≈ 30 USD. GPT-4o five-shot

attains macro-F1 = 0.808 at ≈ 1,750 USD per quarter, two orders of magnitude above the encoder frontier. Med-PaLM 2 zero-shot and GPT-4o zero-shot occupy intermediate positions at macro-F1 of 0.787 and 0.781 respectively. The frontier is dominated by parameter-efficient encoders for stable in-domain workloads.

5. Discussion and Future Work

5.1. Synthesis of Empirical Findings

The evidence from FAERS gold and four supplementary corpora supports a nuanced picture rather than a simple winner. Supervised domain-adapted encoders, particularly PubMedBERT-large, deliver the strongest performance on stable in-domain CTCAE grading at a small fraction of the operating cost of frontier generative models. Parameter-efficient adaptation of medium-scale medical large language models --- represented in our experiments by Me-LLaMA-13B with LoRA --- achieves performance statistically indistinguishable from supervised encoders on the 5-grade task while losing notably less accuracy under domain shift. Zero-shot frontier large language models exhibit competitive recall but weaker precision and calibration; their value lies primarily in scenarios where labeled data is scarce or where rapid prototyping outweighs cost efficiency.

The choice between families hinges on three pragmatic considerations: the stability of the input distribution, the availability of CTCAE-labeled supervision, and the budget envelope for inference. A pipeline operating on a single sponsor's regulatory submissions over years should prefer supervised encoders. A pipeline spanning external real-world data, post-marketing surveillance feeds, and patient-reported channels should consider parameter-efficient generative models. A pipeline with no labeled data and an immediate deployment horizon may legitimately start from zero-shot generative models, accepting precision-recall trade-offs that favor downstream human review.

The persistent gap between automated κ values of 0.715--0.741 and inter-pharmacist agreement of 0.78 suggests that the technical ceiling has not yet been reached, while also indicating that incremental improvements will require advances in CTCAE-aware pretraining objectives or annotated corpora rather than additional scale alone.

5.2. Limitations and Future Directions

Several limitations bound the present findings. The MedDRA-to-CTCAE silver mapping, although validated against a 2,000-report gold subset with substantial inter-rater agreement, retains systematic biases at the grade-2/grade-3 transition where lexical cues are sparse. The gold subset itself reflects two pharmacists' interpretation of CTCAE v5.0, and the upcoming CTCAE v6.0 implementation modifies several grade definitions in ways that will require recalibration. Evaluation was restricted to English text composites, while FAERS receives reports in multiple languages and the international harmonization context demands multilingual support.

The closed-source nature of GPT-4o and Med-PaLM 2 constrains reproducibility; while we recorded snapshot dates and prompt versions, providers can update model weights without notice, and cost figures will date as pricing evolves. Statistical comparisons relied on bootstrap resampling within a single test split rather than multi-site cross-validation, which would offer stronger external validity. We did not perform prompt-injection robustness testing, an increasingly relevant concern when generative models ingest unverified narrative text.

Three directions appear promising. Severity-aware pretraining objectives that explicitly model ordinal CTCAE structure could close part of the residual gap to expert agreement. Multi-center annotation initiatives that share gold CTCAE labels across sponsors and institutions would expand the available supervision beyond what any single research group can produce. Active-learning pipelines that target uncertain regions of the grade space could compress annotation cost while preserving accuracy. We hope the released silver-labeling rules, prompts, and evaluation harness will lower the barrier to

such follow-up work and that future shared tasks will adopt CTCAE 1--5 grading as a first-class evaluation target.

References

1. J. C. Hong, A. T. Fairchild, J. P. Tanksley, M. Palta, and J. D. Tenenbaum, "Natural language processing for abstraction of cancer treatment toxicities: Accuracy versus human experts," *JAMIA Open*, vol. 3, no. 4, pp. 513–517, 2020.
2. S. Chen, M. Guevara, N. Ramirez, A. Murray, J. L. Warner, H. J. W. L. Aerts, T. A. Miller, G. K. Savova, R. H. Mak, and D. S. Bitterman, "Natural language processing to automatically extract the presence and severity of esophagitis in notes of patients undergoing radiotherapy," *JCO Clinical Cancer Informatics*, vol. 7, p. e2300048, 2023.
3. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, and V. Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.
4. C. A. Bejan, M. Wang, S. Venkateswaran, E. A. Bergmann, L. Hiles, and Y. Xu, "irAE-GPT: Leveraging large language models to identify immune-related adverse events in electronic health records and clinical trial datasets," *medRxiv*, 2025.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT 2019*, 2019, pp. 4171–4186.
6. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
7. E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clin. Nat. Lang. Process. Workshop NAACL*, 2019, pp. 72–78.
8. Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, Art. no. 2, 2022.
9. S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. ACL 2020*, 2020, pp. 8342–8360.
10. Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth, X. Zhou, H. He, L. Ohno-Machado, Y. Wu, H. Xu, and J. Bian, "Me-LLaMA: Foundation large language models for medical applications," *npj Digital Medicine*, vol. 8, Art. no. 185, 2025.
11. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings Bioinform.*, vol. 23, no. 6, Art. no. bbac409, 2022.
12. X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu, "A large language model for electronic health records," *npj Digital Medicine*, vol. 5, Art. no. 194, 2022.
13. K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. Aguera y Arcas, and V. Natarajan, "Toward expert-level medical question answering with large language models," *Nature Medicine*, vol. 31, pp. 943–950, 2025.
14. S. Henry, K. Buchan, M. Filannino, A. Stubbs, and Ö. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 1, pp. 3–12, 2020.
15. A. Jagannatha, F. Liu, W. Liu, and H. Yu, "Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0)," *Drug Saf.*, vol. 42, no. 1, pp. 99–111, 2019.
16. S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "Cadec: A corpus of adverse drug event annotations," *J. Biomed. Inform.*, vol. 55, pp. 73–81, 2015.
17. H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo, "Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports," *J. Biomed. Inform.*, vol. 45, no. 5, pp. 885–892, 2012.
18. K. Roberts, D. Demner-Fushman, and J. M. Tonning, "Overview of the TAC 2017 adverse reaction extraction from drug labels track," in *Proc. Text Anal. Conf. (TAC) 2017*, National Inst. Standards Technol., 2017.
19. Y. Li, J. Li, J. He, and C. Tao, "AE-GPT: Using large language models to extract adverse events from surveillance reports — A use case with influenza vaccine adverse events," *PLOS ONE*, vol. 19, no. 3, p. e0300919, 2024.
20. M. Sushil, T. Zack, D. Mandair, Z. Zheng, A. Wali, Y.-N. Yu, Y. Quan, D. Lituiev, and A. J. Butte, "A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports," *J. Am. Med. Inform. Assoc.*, vol. 31, no. 10, pp. 2315–2327, 2024.
21. Y. Hu, Q. Chen, J. Du, X. Peng, V. K. Keloth, X. Zuo, Y. Zhou, Z. Li, X. Jiang, Z. Lu, K. Roberts, and H. Xu, "Improving large language models for clinical named entity recognition via prompt engineering," *J. Am. Med. Inform. Assoc.*, vol. 31, no. 9, pp. 1812–1820, 2024.
22. M. Guevara, S. Chen, S. Thomas, T. L. Chaunzwa, I. Franco, B. H. Kann, S. Moningi, J. M. Qian, M. Goldstein, S. Harper, H. J. W. L. Aerts, P. J. Catalano, G. K. Savova, R. H. Mak, and D. S. Bitterman, "Large language models to identify social determinants of health in electronic health records," *npj Digital Medicine*, vol. 7, Art. no. 6, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.