GAS Journal of Engineering and Technology (GASJET)

Volume 3 | Issue 7, 2026 Homepage: https://gaspublishers.com/gasjet-home/ ISSN: 3048-5800

LLM-Ready CVE Severity Triage with CWE-Aware TF-IDF Features, Probability Calibration, and Evidence-Bound Explanations on 2025 NVD Data

Derek Pan

Computer Science, Boston University, Boston, MA, USA

Received: *20.06.2026* | ***Accepted:*** *25.07.2026* | ***Published:*** *27.07.2026*

***Corresponding Author:** Derek Pan

DOI: *10.5281/zenodo.21624242*

Abstract **Original Research Article**

Vulnerability triage requires severity information that is structured enough for automation, calibrated enough for risk-aware routing, and readable enough for analyst review. This study evaluates four-class CVSS v3.1 severity modeling on the NVD CVE-2025 JSON 2.0 feed using English CVE descriptions, CWE mappings, categorical CVSS vector fields, probability calibration, and an LLM-ready, evidence-bound explanation layer. The feed contained 44,920 CVE records; 39,992 records with a selected non-NONE CVSS v3.1 metric supported a chronological 70/15/15 split of 27,994 training, 5,999 validation, and 5,999 test records. Without CVSS vector tokens, the best early model, a CWE-aware LinearSVC over sparse TF-IDF features, achieved 0.642 accuracy and 0.509 macro-F1. Temperature scaling preserved its decisions while reducing expected calibration error from 0.108 to 0.021. In the vector-aware consistency condition, accuracy reached 0.820 and macro-F1 reached 0.699; Ridge regression reached 0.442 MAE and 0.871 R^2. Because the vector categories define the CVSS base score, these gains are interpreted as score-consistency validation rather than independent severity prediction. The constrained explanation layer produced 5,999 held-out rationales with complete inclusion of the required record fields. That coverage is a schema-consistency check, not a user study or a measure of explanation truthfulness. The findings support a human-in-the-loop workflow in which calibrated text-and-CWE models assist early triage, vector-aware models check scoring consistency, and evidence packets make each recommendation reviewable.

Keywords: CVE severity prediction; National Vulnerability Database; Common Weakness Enumeration; probability calibration; evidence-grounded explanations.

Introduction

The Common Vulnerabilities and Exposures ecosystem provides a shared identifier space for publicly disclosed vulnerabilities, and the National Vulnerability Database enriches many CVE records with metadata such as severity scores, weakness mappings, and affected-product information (CVE Program, n.d.; National Institute of Standards and Technology, n.d.-a, n.d.-b). Severity information is commonly expressed through the Common Vulnerability Scoring System, whose base metrics summarize exploitability and impact in a vector string and a numerical score (Forum of Incident Response and Security Teams, 2019, 2023). Weakness information is commonly normalized

through the Common Weakness Enumeration, a community-maintained taxonomy of software and hardware weakness classes (MITRE, 2024). These structured fields are valuable, but analysts still read CVE descriptions to understand whether a vulnerability resembles remote code execution, injection, access-control failure, information disclosure, denial of service, or another operational pattern.

The practical challenge is that CVE descriptions, CWE mappings, and CVSS vectors do not carry the same type of signal. Descriptions are short natural-language summaries with inconsistent terminology. CWE fields provide categorical root-cause evidence, but many records contain broad or missing weakness labels. CVSS vectors encode severity-defining properties, but they may arrive after initial disclosure or may need independent checking. A severity model therefore has two useful modes. In an early mode, it estimates severity from description and CWE evidence before relying on the CVSS vector. In a vector-aware mode, it learns whether text, CWE, and vector fields jointly reproduce the recorded severity and score. The distinction is important because a model that simply consumes baseScore or baseSeverity would only repeat the target. This study excludes those two target fields from all predictive features.

The early mode is useful for intake queues and quality-control checks. Many security teams see a public disclosure or vendor advisory before all enrichment fields have stabilized. A description-only model cannot replace analyst judgment, but it can flag records whose wording and weakness mapping resemble historically severe cases. The vector-aware mode has a different use. It checks whether a machine-learning model can recover the same severity structure from the CVSS vector categories that analysts and CNAs provide. When the vector-aware model improves sharply over text, the improvement quantifies how much severity information is carried by the vector fields rather than by the prose summary alone.

The CVE-2025 feed is also a useful stress test because it mixes highly repeated vulnerability families with more diverse product-specific records. Web application weaknesses, especially cross-site scripting and SQL injection, appear many times with similar wording, while memory-safety, access-control, and vendor-specific advisories have more varied phrasing. A single model must therefore handle both easy lexical regularities and harder boundary cases where the same terms appear under different impact assumptions. This is one reason the evaluation includes both macro-F1 and ordinal MAE: a model that performs well only on the largest class can look accurate while still being weak on rare but important severity levels.

A further concern is label leakage. CVSS baseScore and baseSeverity are often stored next to the vector fields in NVD JSON. A severity classifier that includes either target field as an input would not be useful, and a regression model that sees baseScore would be circular. The experiments therefore separate target variables from evidence variables. The early feature sets contain no vector information. The vector-aware feature set includes attack and impact categories but still excludes the score and qualitative label. This separation keeps the comparison interpretable: the early results measure language and CWE signal, while the vector-aware results measure how much of the recorded scoring structure can be recovered from the vector categories.

Related Work and Research Gap

Text classification has a long history in information retrieval and machine learning. TF-IDF term weighting remains a strong sparse representation for short documents (Salton & Buckley, 1988), and linear classifiers such as support vector machines have been competitive for high-dimensional text categorization (Joachims, 1998; Sebastiani, 2002). Naive Bayes, ridge-based linear models, and stochastic-gradient classifiers provide useful baselines with different bias-variance and calibration behavior (Bottou, 2012; Hoerl & Kennard, 1970; McCallum & Nigam, 1998; Rifkin & Klautau, 2004). Evaluation requires more than accuracy because severity classes are imbalanced and ordinal: confusing medium with high is less severe than confusing low with critical. We therefore

report accuracy, macro-F1, weighted-F1, ordinal MAE, probability losses, and score-regression metrics, following established evaluation practice for classification, imbalanced data, and probabilistic forecasts (Brier, 1950; H. He & Garcia, 2009; Powers, 2011).

Recent cybersecurity research spans traffic flows, event logs, cloud audit trails, and host traces. Language-guided feature selection has been evaluated for DDoS and intrusion detection (Y. Li & Lu, 2025), whereas deep-autoencoder studies target IoT traffic classification and anomaly detection (Xin et al., 2024). Interpretable linear and hidden Markov models recover attack-chain stages from CloudTrail records (Bai et al., 2026). Outside cybersecurity, multi-head attention has also been evaluated for SSD health-state sequence classification (Wen et al., 2025), which is methodologically suggestive but not direct CVE evidence. These inputs are not interchangeable with CVE descriptions, but together they motivate a transparent pipeline that separates representation, prediction, and analyst-facing evidence.

Operational studies carry the same separation into post-detection action. Cost-aware AIOps routing treats parsing, anomaly detection, diagnosis, and summarization as distinct stages (C. Li et al., 2026). The present study applies this modular logic to vulnerability triage, not runtime-failure diagnosis, and it evaluates the prediction, calibration, and explanation stages separately.

The explanation component addresses a separate need: a triage system should not only return a label but also expose the evidence used for the decision. Transformer encoders and large language models have changed expectations for natural-language interfaces (Brown et al., 2020; Devlin et al., 2019), while local explanation methods emphasize traceability to model inputs (Lundberg & Lee, 2017; Ribeiro et al., 2016). The present study does not train or benchmark a generative LLM. Instead, it implements an LLM-ready, schema-constrained explanation layer that receives only the prediction, calibrated confidence, primary CWE, selected CVSS fields, and short textual cues. This bounded design prevents the explanation from introducing vulnerability facts that are absent from the record.

Evidence-grounded generation offers a second comparator. Evidence-chain RAG audits source attribution (Jin, 2025b), evidence-calibrated RAG permits abstention when retrieval coverage is inadequate (Zhong et al., 2025), and scientific-claim verification ranks support before bounded narration (Su, Chen, et al., 2026). Offline reranking similarly exposes query reformulation, retrieval, and ranking as separable stages (Meng et al., 2026). Independent modeling of RAG retrieval quality reinforces the need to evaluate retrieval separately from generation (R. Zhang et al., 2025). Although no retrieval generator is trained here, these studies support the requirement that each CVE explanation be reconstructable from its evidence packet.

Human-facing presentation is treated as a design implication, not a validated outcome. Scientific-search evidence cards organize ranking, provenance, and uncertainty cues (Jin, 2025c), whereas therapist-facing copilots retrieve and rank support from multi-turn dialogue (Y. Zhang & H. Zhang, 2025a). The same principle motivates a compact ordering of the predicted label, confidence, CWE mapping, CVSS fields, and textual cues in the present study, but this transfer still requires a user study with vulnerability analysts.

Study Positioning and Contributions

This study contributes a reproducible empirical severity-triage benchmark on the CVE-2025 NVD feed. It parses the native JSON 2.0 structure, builds three evidence views, compares seven classifier-feature and four regressor-feature configurations, calibrates the strongest early classifier, and audits held-out explanations for required-field consistency. The analysis asks how much CWE tokens improve early description-based classification, how much CVSS vector evidence improves score-consistency reconstruction, and how calibrated evidence packets can support concise analyst-facing summaries without claiming autonomous severity assignment. The compact-model emphasis is also consistent with cross-domain work that distills quality predictors for edge deployment (B. Zhou et al., 2025), although the

present study does not perform distillation.

The study is deliberately centered on a single annual feed rather than a hand-curated benchmark. That choice keeps the evaluation close to the data format encountered by vulnerability-management tools: brief descriptions, inconsistent weakness granularity, several vulnerability statuses, and multiple scoring sources in one JSON object. It also makes the limitations visible. If a weakness family dominates a month of disclosures, the chronological test set will reflect that skew; if a CNA uses broad wording, the text model must work with that ambiguity. The resulting benchmark is therefore smaller than many natural-language corpora but more representative of day-to-day CVE triage.

Materials and Methods

Data Source and Record Selection

The data source was the NVD CVE-2025 JSON 2.0 feed. The feed is suitable for local experimentation because the compressed file is small relative to typical machine-learning corpora, yet it contains tens of thousands of records and follows the same high-level objects as the NVD 2.0 API response: resultsPerPage, startIndex, totalResults, and vulnerabilities (National Institute of Standards and Technology, n.d.-a, n.d.-b). The parser extracted the English description, publication timestamp, vulnerability status, weaknesses, and metrics for each record. When multiple CVSS v3.1 metrics were present, the canonical metric was chosen by preferring an NVD primary metric, then any primary metric, and then the first listed v3.1 metric. The chosen baseScore and baseSeverity were used as targets, not as features.

Weakness extraction used every weakness description that matched the CWE identifier pattern, while preserving a primary CWE preference for NVD or primary weakness entries. Records without a numeric weakness mapping were assigned the normalized token NVD-CWE-noinfo. Multi-CWE records retained all CWE tokens in the feature representation, but the first preferred CWE was used for summary tables and explanation text. This design avoids discarding secondary weakness evidence while still giving the explanation layer a single compact weakness label.

The parser also retained vulnerability status rather than filtering to only analyzed records. This keeps the experiment close to an operational feed where deferred, modified, awaiting-analysis, and received records may coexist with analyzed records. The target is the selected CVSS v3.1 metric, so records without such a metric do not enter supervised modeling. The procedure yields a training table that is internally consistent: every row has a description, a severity label, a numerical score, at least one weakness token, and the eight CVSS v3.1 base vector fields used in the vector-aware condition.

Severity Labels and Temporal Partitioning

The study uses four actionable severity labels: low, medium, high, and critical. CVSS v3.x defines low as 0.1-3.9, medium as 4.0-6.9, high as 7.0-8.9, and critical as 9.0-10.0 (Forum of Incident Response and Security Teams, 2019). One canonical CVSS v3.1 record had severity none and was excluded from the four-class classifier because it did not provide enough support for a meaningful class. The chronological split sorts records by publication time and assigns the earliest 70% to training, the next 15% to validation, and the latest 15% to testing. This split is stricter than a random split because it asks models trained on earlier disclosures to predict later disclosures.

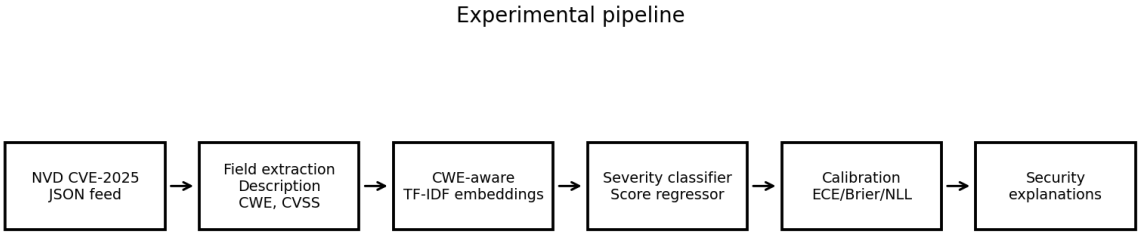

Figure 1. Experimental pipeline from NVD feed parsing to calibrated severity prediction and evidence-bound explanations.

Feature Construction

Three feature configurations were evaluated. The first, Text, uses only the English CVE description. The second, Text+CWE, appends normalized tokens such as cwe_CWE_89 and primary_CWE_89 to the description. This converts weakness mappings into the same sparse feature space as lexical evidence. The third, Text+CWE+CVSS, adds categorical CVSS vector tokens such as attackVector_network, privilegesRequired_low, and confidentialityImpact_high. The vector-aware setting deliberately excludes baseScore and baseSeverity, so the model has to learn the mapping from vector categories rather than copy the target. All text feature matrices used unigram TF-IDF with sublinear term frequency, English stop-word removal, minimum document frequency 10, and a maximum of 15,000 features. These parameters gave stable sparse matrices while preserving the major security terms in the feed (Pedregosa et al., 2011; Salton & Buckley, 1988).

Unigrams were selected over bigrams because the 2025 feed contains many repeated product names, path fragments, and advisory boilerplate. Bigrams increased dimensionality and training time without changing the core scientific comparison between text, CWE-aware text, and vector-aware evidence. The resulting representation is intentionally simple: each model sees a sparse vector whose dimensions correspond to words, normalized CWE markers, or CVSS category tokens. This simplicity makes coefficient inspection and reproduction straightforward, and it avoids relying on pre-trained external embeddings whose training data and update state could differ across environments.

Predictive Models

The classification comparison covers a majority-class baseline, Complement Naive Bayes, RidgeClassifier, LinearSVC, a modified-Huber SGD classifier, and a vector-aware LinearSVC. Complement Naive Bayes represents a generative text baseline (McCallum & Nigam, 1998). LinearSVC and RidgeClassifier represent strong linear discriminative baselines for sparse text (Hoerl & Kennard, 1970; Joachims, 1998; Rifkin & Klautau, 2004). The SGD model gives a fast online-style alternative (Bottou, 2012). The regression configurations predict the numerical CVSS v3.1 base score with Ridge and SGD regression (Bottou, 2012; Hoerl & Kennard, 1970). Predicted scores were clipped to [0, 10] before computing MAE, RMSE, R^2, and severity agreement after applying CVSS thresholds.

Probability Calibration and Evaluation

Calibration was applied to the best early classifier, LinearSVC Text+CWE, because early triage is the setting in which confidence is most useful. Raw LinearSVC decision scores were converted to

probabilities with a softmax transform, then calibrated using three validation-set methods: scalar temperature scaling, multinomial Platt scaling, and one-vs-rest isotonic regression. These methods represent established approaches to mapping classifier scores to usable probabilities (Brier, 1950; Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Platt, 1999). Temperature scaling was selected for downstream per-class analysis because it improved calibration while preserving the LinearSVC ranking and decisions. Selective-decision work in other high-stakes matching settings likewise shows why calibrated probabilities should be paired with abstention or manual review rather than treated as self-justifying scores (Jin, 2025a).

Classification metrics were computed on the held-out test set only. Accuracy measures overall correctness, macro-F1 gives each severity class equal weight, weighted-F1 reflects the empirical class distribution, and ordinal MAE measures the average distance between predicted and reference severity levels under the order low, medium, high, critical. Probability metrics include negative log likelihood, multiclass Brier score, and 15-bin expected calibration error. Regression metrics include MAE, RMSE, R^2, and the severity accuracy obtained by converting each predicted score back into a CVSS severity interval. These metrics jointly capture ranking, class balance, confidence quality, and score fidelity.

Evidence-Bound Explanation Protocol

The explanation layer renders one rationale for every held-out test record. Each evidence packet contains the predicted severity, calibrated confidence, Ridge Text+CWE score estimate, primary CWE, attack vector, privileges required, user interaction, and nonzero confidentiality, integrity, and availability impacts. A short cue extractor selects up to three vulnerability terms from the description, such as injection, overflow, authentication, bypass, disclosure, or denial of service. A schema-constrained template then produces compact prose from those fields. Quality control checks exact inclusion of the CVE identifier, predicted severity, primary CWE, and supplied CVSS fields. This protocol measures field consistency; it does not measure semantic faithfulness beyond the packet, analyst comprehension, trust, or review speed.

Trust also depends on data access and provenance. Federated preference learning demonstrates one way to separate personalization from centralized disclosure (Kuo et al., 2025); accounting-aware retrieval and due-diligence workflows keep claims tied to governed records (Chen et al., 2025); and prompt-injection risk cards make source trust and data-integrity hazards visible to reviewers (Su, Rao, et al., 2026). Accordingly, the present explanation layer exposes only fields derived from the NVD record and does not invent unobserved context.

Reproducibility and Leakage Controls

Implementation used a fixed random seed of 42 and a publication-time split derived directly from the feed timestamps. The same experiment pipeline creates the processed modeling table, trained estimators, calibration bins, figures, and explanation outputs. This design kept the reported numbers stable across repeated local runs and made the relationship between data parsing, feature construction, modeling, and analysis explicit.

Several design choices were made to prevent the evaluation from becoming a duplicate of the NVD severity label. The model never receives baseScore, baseSeverity, or any string formed from those targets as a feature. CVSS vector categories are allowed only in the vector-aware condition, where the goal is to test score consistency rather than early prediction. The early condition is therefore the stricter triage setting: it must infer severity patterns from prose and CWE mappings without directly observing the scoring vector.

Results and Discussion

Dataset Composition and Temporal Split

The feed contained 44,920 CVE records and had format NVD_CVE version 2.0. CVSS v3.1 metrics were available for 39,993 records. After selecting

one canonical v3.1 metric per record and excluding the single none-severity record, the modeling table contained 39,992 records. The data span from 2025-01-01 to 2026-07-06 because CVE-year identifiers can be published after the calendar year in which the identifier was reserved. Table 1 summarizes the chronological split used for all models.

Table 1. Chronological split summary.

Split	Records	First published	Last published	Mean description words	Median base score	Unique primary CWEs
test	5999	2025-12-16 09:16:01.190	2026-07-06 21:16:52.653	49.3	7.1	293
train	27994	2025-01-01 14:15:23.590	2025-10-22 15:15:40.570	54.5	7.1	469
valid	5999	2025-10-22 15:15:40.713	2025-12-16 09:16:01.060	49.7	6.5	296

Table 2. CVSS v3.1 severity distribution by split.

Split	Low	Medium	High	Critical
test	135	2658	2422	784
train	701	12926	10067	4300
valid	144	3072	2033	750

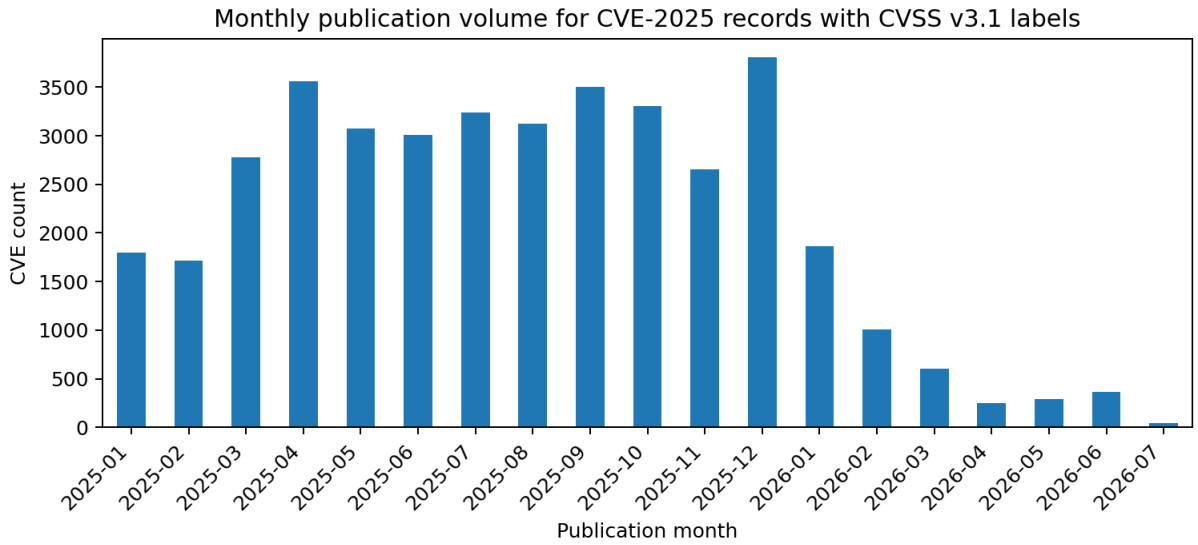

Figure 2. Monthly publication volume after canonical CVSS v3.1 selection.

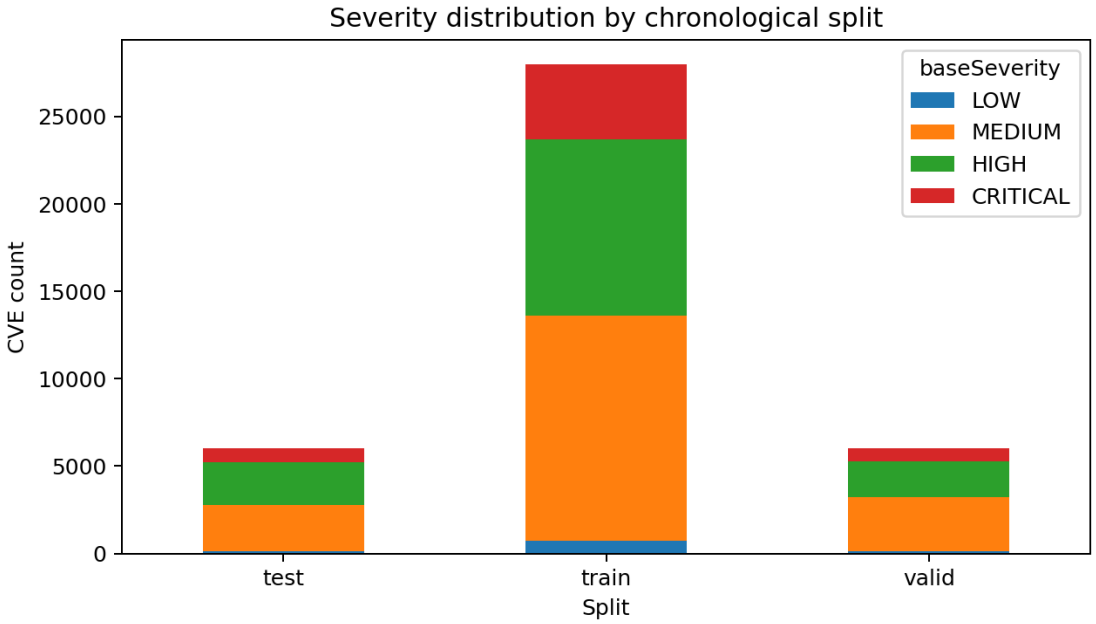

Figure 3. Severity distribution by chronological split.

The split shows a heavy concentration in medium and high severities, with critical records also common and low records relatively scarce. The test set contains only 135 low records, which later explains the instability of low-class recall and precision. The monthly distribution is uneven, with peaks in late 2025 and a tail of CVE-2025 records published in 2026. This temporal structure supports

chronological evaluation because the vocabulary and affected software families change over time.

The status distribution of the raw feed also matters. Analyzed records are the largest group, but deferred and modified records account for a large share of the feed. Keeping these records prevents the model from being tuned only to the most polished subset of NVD entries. It also reflects the practical reality that vulnerability intelligence is updated over time and that security teams may need model output before every enrichment field has been harmonized. Because the supervised target comes from the chosen CVSS v3.1 metric, the modeling table still has a consistent label even when the broader record status differs.

The chronological split also changes how the results should be read. A random split would place near-duplicate vulnerability families from the same disclosure campaigns into both training and testing, which can inflate text-model performance. The time-based split instead asks the model to generalize from earlier wording and weakness distributions to later records. This is harder but more realistic for an intake system that must score tomorrow's disclosures using yesterday's labeled history.

Table 3. Dominant CVSS v3.1 vector values in the modeling data.

CVSS field	Dominant value (share)	Second value (share)	Remaining records
attackVector	NETWORK (78.7%)	LOCAL (18.5%)	1121
attackComplexity	LOW (92.1%)	HIGH (7.9%)	0
privilegesRequired	NONE (51.8%)	LOW (40.1%)	3276
userInteraction	NONE (70.4%)	REQUIRED (29.6%)	0
scope	UNCHANGED (74.5%)	CHANGED (25.5%)	0
confidentialityImpact	HIGH (47.1%)	LOW (30.6%)	8910
integrityImpact	HIGH (39.2%)	LOW (33.3%)	10994
availabilityImpact	HIGH (46.9%)	NONE (37.2%)	6355

Table 4. Top primary CWE mappings and severity profile.

Primary CWE	Records	Mean score	Critical share	High/critical share
CWE-79	7373	6.13	1.0%	21.3%
CWE-89	2683	8.71	51.1%	87.8%
CWE-862	2215	5.83	4.0%	22.1%
NVD-CWE-noinfo	2009	6.46	6.0%	38.4%

Primary CWE	Records	Mean score	Critical share	High/critical share
CWE-352	1688	5.99	1.6%	43.0%
CWE-74	1214	8.91	54.2%	90.8%
CWE-22	944	7.19	16.6%	59.0%
CWE-787	939	7.88	18.2%	83.8%
CWE-476	793	5.88	0.1%	17.5%
CWE-125	770	6.73	4.9%	56.9%

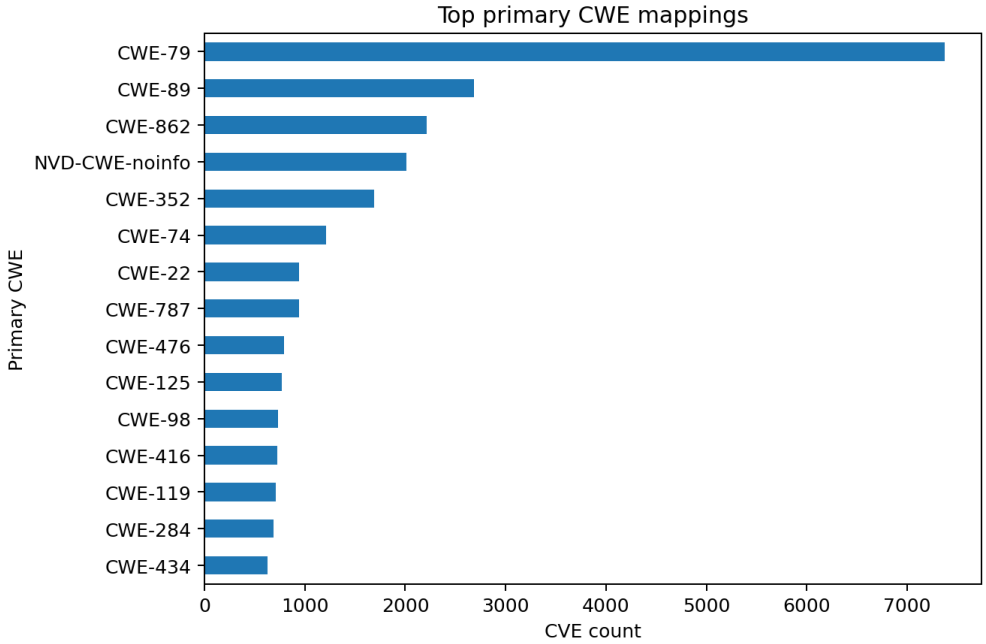

Figure 4. Frequency of the fifteen most common primary CWE mappings.

The vector distribution is operationally plausible. Network attack vector accounts for 78.7% of records, low attack complexity for 92.1%, and no required user interaction for 70.4%. These properties make severity boundaries sensitive to impact and privilege fields. CWE frequency is also concentrated: cross-site scripting (CWE-79) is the most frequent primary mapping, while SQL injection (CWE-89), improper neutralization of special elements (CWE-74), and unrestricted file upload (CWE-434) have much higher high-or-critical rates. This contrast supports a CWE-aware representation because weakness

identifiers carry severity priors that are not fully captured by ordinary lexical terms.

Classification Performance

Table 5. Feature configurations and TF-IDF dimensionality.

Feature set	Description	TF-IDF features	Maximum features	Minimum document frequency	N-gram range
text_only	English description TF-IDF	4658	15000	10	1-1
text_cwe	Description TF-IDF plus CWE tokens	4999	15000	10	1-1
text_cwe_cvss	Description, CWE tokens, and CVSS vector tokens	5021	15000	10	1-1

Table 6. Test-set severity classification results.

Model	Accuracy	Macro-F1	Weighted-F1	Ordinal MAE	ECE
Majority class	0.443	0.154	0.272	0.688	
ComplementNB Text	0.618	0.478	0.613	0.451	0.148
ComplementNB Text+CWE	0.616	0.476	0.609	0.446	0.109
RidgeClassifier Text+CWE	0.596	0.491	0.619	0.507	0.121
LinearSVC Text	0.637	0.506	0.645	0.435	0.110
LinearSVC Text+CWE	0.642	0.509	0.648	0.428	0.108
SGD-modified-huber Text+CWE	0.623	0.480	0.629	0.449	0.144

Model	Accuracy	Macro-F1	Weighted-F1	Ordinal MAE	ECE
LinearSVC Text+CWE+CVSS	0.820	0.699	0.817	0.183	0.178

The classification results show three patterns. First, all learned text models substantially outperform the majority baseline, which reaches only 0.443 accuracy and 0.154 macro-F1. Second, the early Text+CWE LinearSVC is the best non-vector model, reaching 0.642 accuracy and 0.509 macro-F1, a modest gain over text alone. The gain is small because many descriptions already contain weakness terms such as injection, scripting, overflow, bypass, and disclosure. Third, adding CVSS vector tokens changes the task: LinearSVC Text+CWE+CVSS reaches 0.820 accuracy and 0.699 macro-F1. This result is expected because CVSS vector categories define the base-score formula, but the model still has to learn a multi-field nonlinear mapping from sparse categorical tokens.

The difference between weighted-F1 and macro-F1 is important. Weighted-F1 for the early Text+CWE model is 0.648, much higher than macro-F1, because medium and high classes dominate the test set. Macro-F1 exposes the low-class weakness that would be hidden by aggregate accuracy. Ordinal MAE provides another view: the early Text+CWE model has ordinal MAE 0.428, while the vector-aware model reduces it to 0.183. This means vector-aware errors are not only less frequent; they are also closer to the correct severity level when they happen.

The modest CWE gain in the early setting should not be interpreted as weakness information being unimportant. Instead, it indicates that CVE descriptions often repeat weakness names or close paraphrases, so the text channel already contains part of the same evidence. CWE tokens still help stabilize cases where descriptions are short, vendor-specific, or focused on symptoms rather than root cause. They also make the explanation packet clearer because the generated rationale can name the normalized weakness class even when the prose summary uses different wording.

Calibration and Class-Level Error

Table 7. Validation-set calibration methods evaluated on the test set.

Calibration	Accuracy	Macro-F1	NLL	Multiclass Brier	ECE (15 bins)	Mean confidence
Uncalibrated softmax	0.642	0.509	0.918	0.496	0.108	0.534
Temperature scaling T=0.625	0.642	0.509	0.871	0.478	0.021	0.654
Validation Platt	0.664	0.451	0.786	0.446	0.033	0.685

Calibration	Accuracy	Macro-F1	NLL	Multiclass Brier	ECE (15 bins)	Mean confidence
multinomial						
Validation isotonic OVR	0.663	0.456	0.851	0.454	0.018	0.666

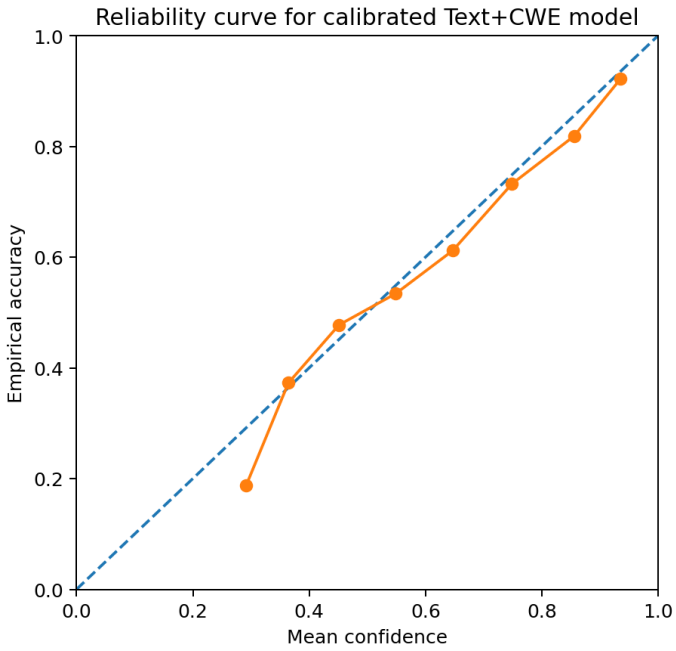

Figure 5. Reliability curve for the temperature-scaled Text+CWE LinearSVC.

Calibration changed the usefulness of probabilities even when the classifier decisions were unchanged. The uncalibrated softmax over LinearSVC scores had ECE 0.108. Temperature scaling with T = 0.625 kept accuracy and macro-F1 fixed while reducing ECE to 0.021 and lowering negative log likelihood from 0.918 to 0.871. Multinomial Platt scaling achieved the lowest negative log likelihood, but it changed decision boundaries and reduced macro-F1 because the small low class and the medium-high boundary were reweighted. For triage explanations, temperature scaling is therefore the most balanced option: it improves confidence quality without changing the model's ranked severity choice.

The reliability curve shows why this matters in practice. Before calibration, the average confidence of the early classifier is lower than its empirical accuracy in several bins, and the predicted probabilities are not easy to interpret. After temperature scaling, the points move closer to the diagonal, so a confidence around 0.7 can be read as

a roughly comparable empirical success rate. This does not make the model certain; it makes the confidence score a more honest input for queues, dashboards, and generated explanations.

Calibration also affects how analysts might act on model output. Without calibration, a confidence threshold such as 0.80 is only a ranking threshold; it does not carry a clear empirical meaning. After calibration, threshold-based routing becomes easier to justify. For example, a queue can separate high-confidence high-or-critical predictions from medium-confidence boundary cases, while still preserving the record-level evidence needed for analyst review. This is especially valuable when the model is used before final enrichment has settled.

Calibrated probabilities still do not define an action policy. Selective-refusal research shows how uncertain matches can be deferred (Jin, 2025a), while calibration-light subject-independent modeling shows that transfer settings still require explicit uncertainty checks (Wu, Mi, et al., 2025). Offline counterfactual evaluation asks a different question: whether an action policy would improve outcomes (Mu et al., 2025). News-based uncertainty forecasting likewise separates directional prediction from decision risk (H. Zhou & Zhang, 2025). In operational security, predictive optimization of DDoS mitigation makes the policy layer explicit (B. Wang et al., 2024). For vulnerability management, remediation thresholds and staffing policies therefore need a separate cost-sensitive evaluation beyond the classifier metrics reported here.

Table 8. Per-class metrics for the temperature-scaled Text+CWE model.

Severity	Precision	Recall	F1	Support
LOW	0.129	0.274	0.176	135
MEDIUM	0.728	0.746	0.737	2658
HIGH	0.676	0.581	0.625	2422
CRITICAL	0.466	0.537	0.499	784

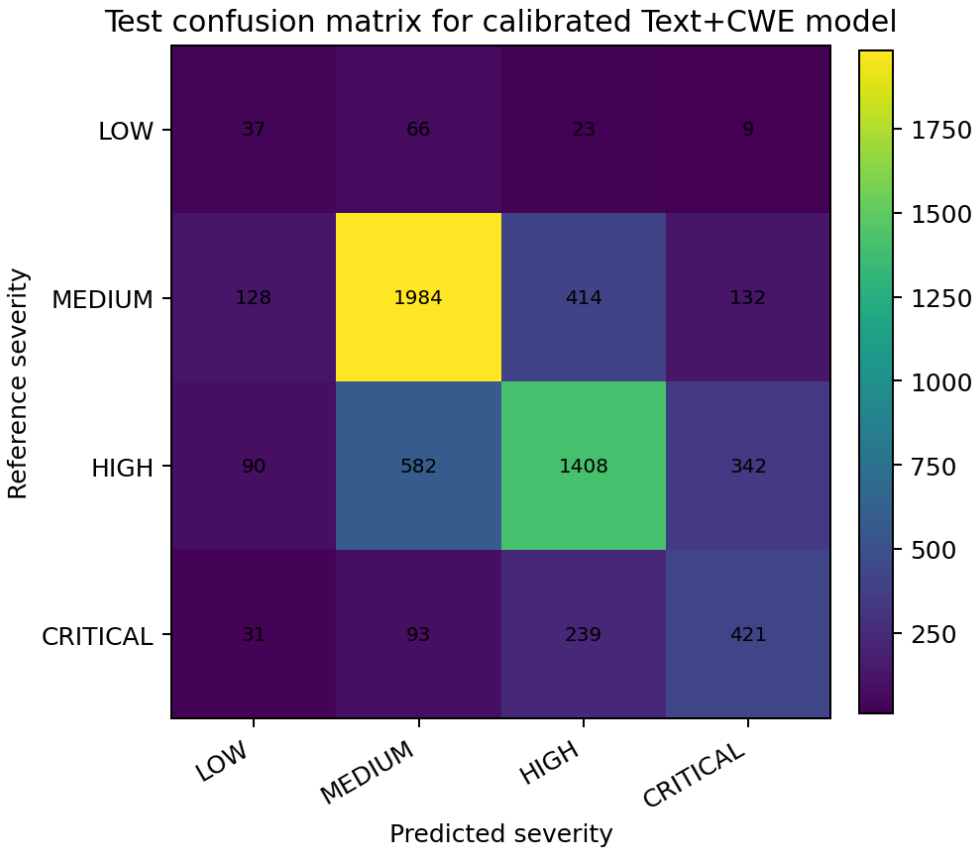

Figure 6. Confusion matrix for the temperature-scaled Text+CWE model.

Per-class results reveal that medium and high records are much easier than low records. Medium has F1 0.737 and high has F1 0.625, while low has F1 0.176 because the test split contains only 135 low records and many low descriptions share vocabulary with medium issues. Critical reaches F1 0.499 and is frequently confused with high. The confusion matrix confirms that most errors occur between adjacent severity levels: low-to-medium, medium-to-high, high-to-medium, and critical-to-high. This is preferable to arbitrary errors because severity is ordinal, but it still matters for patch priority when the threshold between high and critical drives service-level agreements.

Weakness- and Status-Slice Analysis

Table 9. Test accuracy slices for frequent primary CWE mappings.

Primary CWE	Records	Accuracy	Mean score
CWE-79	925	0.853	6.28

Primary CWE	Records	Accuracy	Mean score
CWE-862	489	0.748	5.91
CWE-98	326	0.985	8.04
CWE-89	241	0.589	8.46
NVD-CWE-noinfo	238	0.618	6.62
CWE-22	164	0.445	7.24
CWE-502	145	0.476	8.72
CWE-352	143	0.825	5.74

Table 10. Test accuracy slices by vulnerability status.

Vulnerability status	Records	Accuracy	Mean score
Analyzed	2893	0.577	7.02
Awaiting Analysis	100	0.460	7.36
Deferred	2340	0.724	6.79
Modified	657	0.661	7.54
Received	3	0.000	6.60
Undergoing Analysis	6	1.000	6.70

Slice analysis shows that accuracy varies strongly by weakness family. CWE-98 records are highly regular in the test period and reach 0.985 accuracy, while CWE-22, CWE-89, and CWE-502 are harder because their descriptions cover multiple exploit paths and impact levels. Status slices show a different pattern: deferred records reach 0.724 accuracy, analyzed records 0.577, modified records 0.661, and awaiting-analysis records 0.460. These values do not imply that one status is intrinsically easier; they show that status groups have different mixtures of vendors, weakness families, and severity labels. The slice results are useful for deployment because they identify where analysts should be careful with automated confidence.

Score Regression and Feature Interpretation

Table 11. CVSS v3.1 base-score regression results.

Model	MAE	RMSE	R^2	Severity accuracy from score
Ridge Text	0.969	1.276	0.398	0.642
Ridge Text+CWE	0.957	1.264	0.410	0.636
Ridge Text+CWE+CVSS	0.442	0.592	0.871	0.834
SGDReg Text+CWE	1.130	1.447	0.226	0.598

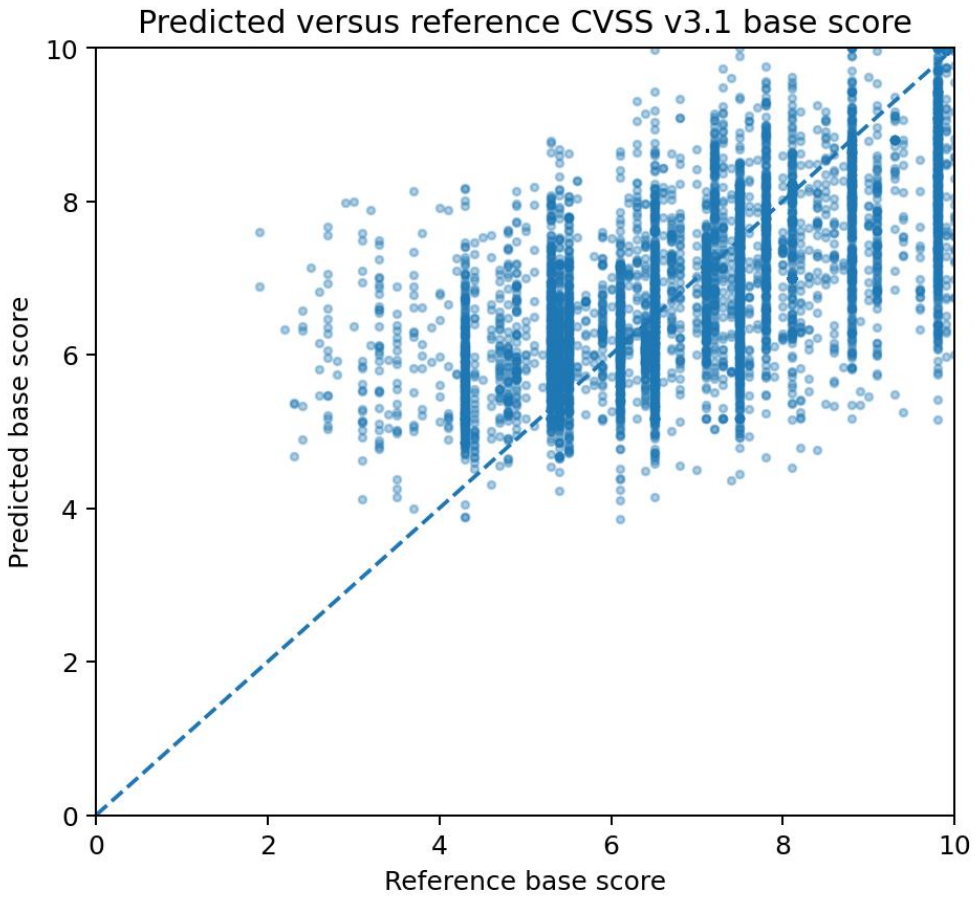

Figure 7. Ridge Text+CWE predicted score versus reference base score on the test set.

Regression results are consistent with the classification results. Ridge Text+CWE has MAE 0.957 and R^2 0.410, indicating that text and CWE evidence provide useful but incomplete estimates of the continuous CVSS score. Ridge Text+CWE+CVSS has MAE 0.442 and R^2 0.871, demonstrating that vector categories sharply constrain the score. The scatter plot for the early Ridge model shows regression to common scores near the medium-high boundary, while the vector-aware regression separates high-impact records more clearly. This confirms that early triage and vector-aware verification should be treated as related but different tasks. If a future system must combine several analyst or model rankings rather than reproduce one recorded score, spectral rank-aggregation guarantees offer a distinct evaluation problem (Zhong & Ling, 2024a).

The regression view is useful because many operational policies are score-based rather than label-based. Two vulnerabilities can both be high severity while sitting at different points in the 7.0-8.9 interval, and that difference may affect patch sequencing when teams face a large backlog. The early Ridge model is not accurate enough to replace scoring, but it gives a continuous estimate that can be presented alongside the calibrated class. The vector-aware Ridge model, by contrast, is best understood as a consistency model: it checks whether the vector categories support the recorded numerical score.

Table 12. Top three positive class-specific cues in the early Text+CWE LinearSVC.

Severity	Rank	Feature	Coefficient
LOW	1	reported	2.78
LOW	2	cuda	2.67
LOW	3	consumption	2.57
MEDIUM	1	forgery.this	3.45
MEDIUM	2	dom-based	2.36
MEDIUM	3	cwe_cwe_79	2.18
HIGH	1	reflected	4.30
HIGH	2	xss.this	3.69
HIGH	3	uaf	2.53
CRITICAL	1	cwe_cwe_74	2.69
CRITICAL	2	server.this	2.31
CRITICAL	3	primary_cwe_434	2.30

Explanation Consistency and Operational Implications

Table 13. Explanation-layer consistency checks on the held-out test set.

Metric	Value
test explanations	5,999
CVE ID coverage	100.0%
predicted label coverage	100.0%
primary CWE coverage	100.0%
CVSS field consistency	100.0%
median words	67
mean words	66.5

The feature-cue table provides a simple sanity check on the early model. CWE-79 is a strong medium cue, while SQL injection and other injection-related terms are more associated with high or critical severities. Some high-weight features are product- or source-specific because CVE descriptions include vendor vocabulary and disclosure conventions. The explanation layer does not expose raw coefficients to analysts. Instead, it translates the calibrated prediction and field evidence into a compact rationale. Across 5,999 test records, every rendered explanation contained the CVE identifier, predicted label, primary CWE, attack vector, privileges required, and user-interaction field, with a median length of 67 words. A typical explanation states the predicted severity, the confidence, the estimated score, the weakness mapping, the attack-vector conditions, and the CIA impact profile. This style is concise enough for triage dashboards and structured enough for automated consistency checks.

Across domains, explanation cards expose evidence hierarchy rather than replace judgment. Recommendation cards pair product rationales with trust cues (B. Zhang et al., 2025); medical interfaces visualize uncertainty and rubric-based safety signals (Ye et al., 2025; C. Li et al., 2025); and financial-risk dashboards examine chart hierarchy and explainability for institutional users (Z. Li et al., 2025). In security operations, distributed-log incident cards organize service context and actionable evidence (Nie et al., 2026). These precedents motivate ordering the predicted label, confidence, CWE, CVSS fields, and textual cues, but they do not establish usability for vulnerability analysts.

Human-centered design work further suggests that machine-produced rationales should remain inspectable and editable. Text-grounded rationale cards translate layout metadata into explicit design decisions (Wu, Meng, et al., 2025), visual briefs preserve the path from intent to action (Tu et al., 2025), and designer-editable cards keep structured outputs open to revision (Mu et al., 2026). Review-derived adaptation in a volleyball learning interface provides another domain-specific example (J. Zhang, 2025). Because agreement between AI critique and human judgment must be measured rather than assumed (Y. Li et al., 2025), the CVE explanation is treated as a review artifact, not a final adjudication.

Table 14. Representative held-out explanations generated from calibrated evidence packets.

CVE ID	Evidence packet	Generated explanation
CVE-2025-68055	actual=HIGH; pred=HIGH; confidence=0.61; CWE=CWE-89	The calibrated model assigns high severity to CVE-2025-68055 with 0.61 confidence and an estimated CVSS score of 8.5. The main evidence is sql, injection, command, weakness mapping CWE-89, attack vector NETWORK, privileges required LOW, user interaction NONE, ...
CVE-2025-68056	actual=HIGH; pred=HIGH; confidence=0.89; CWE=CWE-89	The calibrated model assigns high severity to CVE-2025-68056 with 0.89 confidence and an estimated CVSS score of 9.1. The main evidence is sql, injection, command, weakness mapping CWE-89, attack vector NETWORK, privileges required LOW, user interaction NONE, ...
CVE-2025-68061	actual=HIGH; pred=HIGH; confidence=0.91; CWE=CWE-98	The calibrated model assigns high severity to CVE-2025-68061 with 0.91 confidence and an estimated CVSS score of 8.2. The main evidence is remote, file, weakness mapping CWE-98, attack vector NETWORK, privileges required LOW, user interaction NONE, and impact ...

The sample explanations demonstrate the intended behavior. They do not attempt to infer exploit availability, affected deployments, or business impact. They repeat the predicted severity and confidence, name the weakness mapping, and list the vector fields that drive the rationale. This makes the prose useful as a human-readable summary while keeping it close enough to the original evidence for automatic checking. In a production setting, the same evidence packet could be passed to a stronger language model for stylistic variation, but the factual contract should remain unchanged.

Overall, the experiments support a two-layer triage design. A CWE-aware text model provides early estimates when only descriptions and weakness mappings are available. A vector-aware model provides a stronger score-consistency check once CVSS base metrics are present. Calibration is essential when model outputs are used in explanations or prioritization queues, because a severity label without a trustworthy confidence score can lead to overreaction or underreaction. The constrained language layer then communicates the

model result without introducing unsupported details.

The error profile also suggests that the best deployment pattern is human-in-the-loop. Adjacent-class mistakes are common, especially around the medium-high and high-critical boundaries, so model output should be a prioritization signal rather than a final label. The explanations are designed for exactly that workflow: they compress the evidence that the model saw, expose the calibrated confidence, and make it easier for a reviewer to decide whether the prediction is sensible or whether the record deserves manual rescoring.

The practical interpretation is not that machine learning should replace CVSS scoring. Instead, the results show where automation can help. During intake, a calibrated early model can sort records for analyst attention and highlight cases whose text and CWE profile resemble severe vulnerabilities. During enrichment review, the vector-aware model can act as a consistency check between text, CWE, and the CVSS vector. During communication, the explanation layer can summarize why the model produced a label and which fields should be checked by an analyst. These roles are complementary and keep the final responsibility with security teams.

The model behavior also suggests useful operational thresholds. High-confidence correct examples often come from repeated weakness families with consistent descriptions, such as SQL injection or PHP file inclusion. Lower-confidence examples often occur near the medium-high boundary or in families where the same weakness can have different impacts depending on privileges, authentication, or data exposure. A triage deployment can route high-confidence high or critical predictions to accelerated review while sending low-confidence or boundary predictions to manual scoring. In that sense, calibration is not an academic add-on; it determines whether confidence can be converted into workflow rules.

Limitations and Future Work

The primary limitation is that the target labels are the recorded CVSS v3.1 severities, not independent measurements of exploitability in a particular deployment. CVSS is a standardized severity framework, but it is not a complete risk model and does not include local asset value, exposure, compensating controls, or active exploitation in the base score (Forum of Incident Response and Security Teams, 2019). The model therefore predicts recorded severity, not enterprise risk.

The second limitation is temporal scope. The experiment uses CVE-2025 identifiers from one NVD feed snapshot, and the test period includes records published after December 2025. This is correct for a CVE-year feed, but it means the analysis is a snapshot study rather than a live monitoring system. Future work should repeat the same protocol across multiple CVE years and measure how quickly models degrade as products, vulnerability types, and scoring practices change.

The third limitation concerns textual cues. Descriptions sometimes contain explicit severity words such as critical or high. The study keeps those words because they are part of the operational record available to a triage system, but a stricter early-warning setting could mask them. The low class also has limited support in the chronological test split, which makes low-severity performance less stable than medium or high performance.

The fourth limitation is that the explanation layer is evidence-bound. This design reduces unsupported claims, but it also limits explanatory richness. The generated rationale does not inspect vendor advisories, proof-of-concept repositories, exploit telemetry, or organization-specific exposure. It should therefore be treated as a triage explanation rather than a final security assessment. The present field-coverage check should therefore not be described as a usability or trust evaluation.

A final limitation is that the study uses compact linear models rather than deep encoders. This choice makes the full-feed experiment fast and transparent, but learned representations are a plausible extension whose transfer evidence must be interpreted cautiously. Comparative short-text work distinguishes encoder and decoder behavior (Xu et al., 2025); cross-modal pretraining enables few-shot transfer on MedMNIST (Mi et al., 2025); and

medical vision-language studies evaluate robustness under modality shift (S. Lu & Zou, 2026). Attention-enhanced object detection in autonomous-driving imagery provides another architecture-transfer example (Ling et al., 2024), but it supplies no direct evidence for vulnerability text. Cold-start forecasting for new inference tenants likewise emphasizes evaluation under sparse history (S. He et al., 2025). These studies motivate controlled transformer baselines; they do not establish gains on temporally split CVE records.

Operational deployment introduces constraints not measured in this benchmark. Multi-horizon GPU forecasting turns workload semantics into capacity risk curves (Zhao et al., 2025). Task-level GPU-memory prediction focuses more narrowly on model resource requirements (C. Wang et al., 2025), while power-aware inventory planning combines job-level forecasts with workload explanations (S. He et al., 2026). A deployed CVE triage service should therefore report inference cost and queue delay in addition to accuracy and calibration.

Resource forecasts still need governance and presentation layers. Constrained budgeting work integrates demand forecasts with response models instead of treating prediction as an allocation rule (Y. Lu et al., 2025). Progressive document rendering makes time-to-functional display an explicit objective (H. Wang et al., 2025), while a DevOps copilot study evaluates retrieval-augmented runbooks together with command safety (B. Zhang et al., 2026). Analogously, CVE triage should measure time to a reviewable evidence card and verify any recommended action before execution.

Evidence-linked support is also role and domain dependent. Therapist-facing recommendation cards expose the evidence behind counseling support (Y. Zhang & H. Zhang, 2025b). In a different setting, teacher-facing early-warning systems pair educational predictions with explanations (J. Zhang, 2026). These analogies reinforce the need for role-specific evaluation: future work should test the CVE card with vulnerability analysts instead of inferring usefulness from field coverage.

Three failure modes deserve targeted follow-up. Numerical-reasoning guardrails show why generated summaries should be checked against source values (Z. Li et al., 2026); multilingual RAG studies show that calibration can vary across language and access conditions (Chen & Xu, 2026); and estimator-specific uncertainty theory cautions against treating one calibration result as universal (Zhong & Ling, 2024b). Evidence-constrained monitoring of governed disclosure and settlement records adds a provenance perspective (S. Zhou et al., 2026). The analogous CVE audit should therefore test numeric fidelity, multilingual descriptions, and stage-specific failure propagation, not only the presence of required fields. It should also mask explicit severity words, deduplicate near-identical disclosure families across splits, archive the exact NVD snapshot timestamp, and report package and hardware versions.

Integrated Interpretation

The integrated results support a bounded, two-stage triage architecture. From 44,920 CVE-2025 feed records, 39,992 records supported four-class CVSS v3.1 modeling. The early Text+CWE LinearSVC achieved 0.642 accuracy and 0.509 macro-F1, and temperature scaling reduced ECE from 0.108 to 0.021 without changing its decisions. Adding CVSS vector categories raised classification accuracy to 0.820 and reduced score-regression MAE to 0.442, but that condition is best understood as score-consistency validation because those categories define the base score. The schema-constrained explanation layer preserved required-field coverage for all 5,999 test records; it did not establish explanation truthfulness or analyst trust. Accordingly, the defensible use is human-in-the-loop: description and CWE features can prioritize early review, vector-aware models can flag scoring inconsistencies after enrichment, and evidence packets can make the basis of each prediction inspectable.

References

Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing.

Informatics, Electrical and Electronics Engineering (Infotron), *6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, & K.-R. Muller (Eds.), *Neural networks: Tricks of the trade* (2nd ed., pp. 421–436). Springer. https://doi.org/10.1007/978-3-642-35289-8_25

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design*, *4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering*, *4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

CVE Program. (n.d.). *Common Vulnerabilities and Exposures.* Retrieved July 23, 2026, from https://www.cve.org/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Forum of Incident Response and Security Teams. (2019). *Common Vulnerability Scoring System v3.1: Specification document.* https://www.first.org/cvss/v3.1/specification-document

Forum of Incident Response and Security Teams. (2023). *Common Vulnerability Scoring System v4.0: Specification document.* https://www.first.org/cvss/v4.0/specification-document

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, *70* (pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering*, *4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering*, *5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*,

12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering*, *4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

Jin, J. (2025b). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering*, *4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

Jin, J. (2025c). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design*, *3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Machine learning: ECML-98* (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683

Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering*, *4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering*, *5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design*, *3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering*, *4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design*, *3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering*, *5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design*, *3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

Ling, Z., Xin, Q., Lin, Y., Su, G., & Shui, Z. (2024). Optimization of autonomous driving image detection based on RFAConv and triplet attention. *Applied and Computational Engineering*, *77*(1), 210–217. https://doi.org/10.54254/2755-2721/77/2024MA0067

Lu, S., & Zou, T. (2026). Uncertainty-aware medical vision–language classification on a lightweight MedMNIST-compatible biomedical patch benchmark. *Journal of Technology Informatics and Engineering*, *5*(2), 1–19. https://doi.org/10.51903/jtie.v5i2.530

Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained,

data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. *Journal of Technology Informatics and Engineering*, *4*(3), 493–520. https://doi.org/10.51903/jtie.v4i3.466

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (pp. 41–48). AAAI Press.

Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering*, *5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering*, *4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

MITRE. (2024). *Common Weakness Enumeration list (Version 4.20).* https://cwe.mitre.org/data/index.html

Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. *International Journal of Graphic Design*, *4*(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering*, *4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

National Institute of Standards and Technology. (n.d.-a). *National Vulnerability Database.* Retrieved July 23, 2026, from https://nvd.nist.gov/

National Institute of Standards and Technology. (n.d.-b). *NVD CVE API 2.0: Vulnerability APIs.* Retrieved July 23, 2026, from https://nvd.nist.gov/developers/vulnerabilities

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). Association for Computing Machinery. https://doi.org/10.1145/1102351.1102430

Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design*, *4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. L. Bartlett, B. Scholkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://arxiv.org/abs/2010.16061

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, *5*, 101–141. https://jmlr.org/papers/v5/rifkin04a.html

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering*, *5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design*, *4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. *International Journal of Graphic Design*, *3*(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. *Applied and Computational Engineering*, *64*(1), 89–94. https://doi.org/10.54254/2755-2721/64/20241350

Wang, C., Wen, Z., Zhang, R., Xu, P., & Jiang, Y. (2025). GPU memory requirement prediction for deep learning task based on bidirectional gated recurrent unit optimization transformer. In *2025 5th International Conference on Artificial Intelligence, Virtual Reality and Visualization (AIVRV)*. IEEE. https://doi.org/10.1109/AIVRV67401.2025.11350369

Wang, H., Ren, Y., & Chang, X. (2025). Layout-aware progressive PDF rendering: AI prioritization of PDF slices to reduce time-to-functional-first-frame on FUNSD. *Journal of Technology Informatics and Engineering*, *4*(2), 425–446. https://doi.org/10.51903/jtie.v4i2.523

Wen, Z., Zhang, R., & Wang, C. (2025). Optimization of bi-directional gated loop cell based on multi-head attention mechanism for SSD health state classification model. In *2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)* (pp. 548–552). IEEE. https://doi.org/10.1109/ICECAI66283.2025.11171441

Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design*, *3*(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

Wu, Q., Mi, G., & Wood, D. (2025). Calibration-light subject-independent motor imagery BCI

via self-supervised pretraining and Conformer. *Journal of Technology Informatics and Engineering*, *4*(1), 239–262. https://doi.org/10.51903/jtie.v5i1.493

Xin, Q., Xu, Z., Guo, L., Zhao, F., & Wu, B. (2024). IoT traffic classification and anomaly detection method based on deep autoencoders. *Applied and Computational Engineering*, *69*(1), 64–70. https://doi.org/10.54254/2755-2721/69/20241511

Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering*, *4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224. *International Journal of Graphic Design*, *3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design*, *3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering*, *5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

Zhang, J. (2025). Adaptive user interface design for volleyball learning apps: Empirical evidence from Google Play reviews and mobile screen analysis. *International Journal of Graphic Design*, *3*(1), 175–195. https://doi.org/10.51903/ijgd.v3i1.3618

Zhang, J. (2026). Early warning, grade prediction, and teacher-facing LLM-ready explanations toward an open volleyball course: Reproducible evidence from four public education datasets. *Journal of Technology Informatics and Engineering*, *5*(2), 20–44. https://doi.org/10.51903/jtie.v5i2.525

Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). *Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.19481

Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering*, *4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design*, *3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba clusterdata GPU trace. *Journal of Technology Informatics and Engineering*, *4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering*, *4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA*, *13*(3), Article iaae020. https://doi.org/10.1093/imaiai/iaae020

Zhong, Z. S., & Ling, S. (2024b). *Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.05944

Zhou, B., Wang, H., & Chang, X. (2025). Distilling VMAF into an edge-deployable quality predictor: A pilot shot-level proxy with LLM-ready quality tokens. *Journal of Technology Informatics and Engineering*, *4*(2), 447–463. https://doi.org/10.51903/jtie.v4i2.522

Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from the 2015–2024 FRED panel. *Journal of Technology Informatics and Engineering*, *4*(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540

Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering*, *5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544