Stout in Environmental Studies (SES)

DOI: https://doi.org/10.61424/p86cmv54
Journal Homepage: https://stoutjournals.org/index.php/SES

Energy-Efficient Narrative-Aware Climate Claim Verification with Calibrated Evidence Ranking and Selective Abstention

Xavier Xu
Computer Science, University of Florida, Gainesville, FL, USA
Corresponding Author: Xavier Xu, **E-mail**: x.xu_cs@protonmail.com

ARTICLE INFORMATION

Submitted: December 28, 2025
Accepted: May 13, 2026
Published: August 01, 2026
Volume: 1
Issue: 1

KEYWORDS

Climate misinformation; scientific claim verification; evidence reranking; narrative-aware modeling; probability calibration; selective abstention; energy-efficient natural language processing.

ABSTRACT

Climate claims circulating in public discourse often compress complex scientific findings into categorical statements, making evidence selection and uncertainty communication central to reliable automated verification. This study evaluates a resource-conscious pipeline on ClimateCheck 2025, comprising 3,048 human-annotated claim–abstract pairs. The experiments address closed-pool evidence reranking and three-way classification as Supports, Refutes, or Not Enough Information. Comparators include BM25, TF–IDF cosine similarity, latent semantic analysis, a quantized 135M-parameter causal language model, a character n-gram classifier, sparse logistic regression, and lexical–semantic models with and without cross-fitted predicted narrative probabilities. Claim-grouped cross-fitting prevents identifier leakage, while temperature scaling and training-locked abstention screening address uncertainty. On 1,904 held-out pairs, Word+Char TF–IDF Logistic led ranking with mean average precision 0.670; LSA-128, the narrative-aware reranker, BM25, and the compact causal-LM hybrid obtained 0.665, 0.657, 0.649, and 0.647, respectively. The narrative-aware verifier produced the highest macro-F1, 0.394, compared with 0.392 for Word+Char TF–IDF Logistic, although the paired difference was not significant after Holm adjustment. Narrative-conditioned calibration reduced adaptive expected calibration error from 0.035 under global scaling to 0.025 and produced an area under the risk–coverage curve of 0.551. Neither prespecified target-risk screen yielded a feasible abstention threshold. Warm narrative-aware inference used 2.798 CPU seconds per 1,000 pairs, corresponding to a standardized 15 W CPU-time proxy of 1.166×10^{-5} kWh. The results establish a measured local benchmark in which ranking, verification, probability quality, and computational cost remain separately interpretable.

1. Introduction

Climate communication presents an unusual verification problem. Short categorical claims contrast with the scientific record's cautious language, study-specific scope, and qualified conclusions. Repeated misinformation narratives can influence how people interpret subsequent evidence, and interventions are more effective when they expose misleading argument structures rather than merely repeat a correction (Cook et al., 2017). Automated systems must therefore identify relevant science, distinguish support from contradiction and insufficiency, express uncertainty faithfully, and remain computationally practical.

Evidence-grounded fact checking is commonly organized as a retrieve-then-verify pipeline. FEVER established a large-scale formulation in which evidence selection and label prediction are evaluated together (Thorne et al., 2018), while SciFact adapted this logic to scientific claims and research abstracts (Wadden et al., 2020). Climate-specific resources introduced additional domain constraints. CLIMATE-FEVER connected real-world climate claims to evidence passages (Diggelmann et al., 2020), and ClimateCheck linked 435 lay climate claims to 3,048 scholarly claim–abstract pairs annotated as Supports, Refutes, or Not Enough Information (Abu Ahmad et al., 2025a). The ClimateCheck shared task separately evaluated abstract retrieval and evidence-conditioned verification, encouraging multi-stage systems that combine lexical retrieval, learned reranking, and language-model analysis (Abu Ahmad et al., 2025b).

Recent research broadens retrieve–verify systems from relevance scoring toward evidence-chain completeness, provenance, and calibrated handling of unresolved questions. Word-level and deterministic evidence-chain frameworks separate source attribution from fluent answer generation (Jin, 2025a; C. Li et al., 2024), while unanswerable-question work makes retrieval coverage and abstention jointly observable (Zhong, Chen, et al., 2025). Budgeted multi-hop retrieval and retrieval-quality prediction add explicit questions about whether another search step is useful and whether the resulting candidate set is strong enough to support a decision (Su et al., 2025; R. Zhang et al., 2025). A ClimateCheck-specific agent further connects evidence ranking to narrative-aware verification (Su, Chen, et al., 2026). These studies motivate the present separation of ranking, relation prediction, and abstention; they do not change the single-abstract, closed-pool boundary of the experiments.

High-capacity architectures can improve semantic matching, but their fitting and inference costs also matter. Large-scale NLP has made computation a substantive evaluation dimension (Strubell et al., 2019). Transparent reporting requires a defined hardware boundary, runtime, measurement source, and normalization unit (Henderson et al., 2020), especially when a fact-checking workflow scores many pairs before human review.

Resource-aware evaluation is also supported by adjacent operational research. Capacity forecasting, workload-semantic demand modeling, and GPU-memory prediction show why an AI service must be interpreted against an explicit demand and hardware boundary (He et al., 2023; C. Wang et al., 2025; S. Zhao et al., 2025). Hybrid-cloud placement and profit-aware admission control add deployment-cost and service-risk constraints (Xin, 2025b; S. Zhao et al., 2026). Compact intrusion models, lightweight medical foundation models, and edge-oriented quality distillation illustrate the complementary effort to reduce inference burden at the model level (Lu & Zhou, 2024; Mi et al., 2025; B. Zhou, Wang, et al., 2025). Accordingly, the energy values reported here are local workload measurements, not cross-hardware claims.

The study examines a lightweight, narrative-aware pipeline under a controlled CPU budget. The 2025 data contain annotated claim–abstract pairs, not the separate 394,269-document publications corpus used for the original open-corpus task. Ranking is therefore closed-pool reranking: for each exact claim, the system orders only its annotated candidates. This measures whether evidentiary abstracts outrank Not Enough Information candidates; it does not reproduce the full-corpus leaderboard.

Three research questions organize the study. First, how much do supervised lexical–semantic and predicted-narrative features improve candidate-pool ranking beyond BM25, TF–IDF, and latent semantic analysis? Second, how accurately and reliably can compact models classify the three relations? Third, what trade-offs emerge among quality, calibration, selective coverage, and energy per 1,000 pairs?

The study contributes a version-controlled data audit; local ranking and verification comparisons, including a quantized 135M-parameter causal-LM reranker; and a leakage-safe narrative signal learned from claim text while gold narrative codes remain restricted to subgroup analysis. It joins these elements with claim-clustered intervals, post-hoc calibration, training-locked abstention screening, and a standardized CPU-time energy proxy.

2. Literature Review

Climate claim verification sits at the intersection of misinformation research, scientific document retrieval, and uncertainty-aware classification. General fact-checking datasets capture broad claim variation: MultiFC assembled claims from multiple real-world fact-checking organizations and domains (Augenstein et al., 2019), whereas FEVER emphasized retrieval of evidence sufficient to support, refute, or leave a claim unresolved (Thorne et al., 2018). SciFact moved the evidence source to scientific abstracts and included sentence-level rationales, establishing a close methodological precedent for claim–research alignment (Wadden et al., 2020). HoVer further showed that some propositions require evidence composition across multiple documents, highlighting a limitation of systems that classify each pair in isolation (Jiang et al., 2020). Retrieval-augmented generation extends the same retrieve-then-reason architecture by coupling a neural retriever to a generator (Lewis et al., 2020). For multi-hop verification, this perspective makes evidence-chain completeness and provenance explicit concerns, even when a pairwise experiment cannot measure them.

Climate-focused resources narrow the domain while increasing the importance of scientific provenance. CLIMATE-FEVER contains real-world climate claims paired with evidence, providing a bridge between generic fact checking and domain-specific verification (Diggelmann et al., 2020). ClimateCheck differs by mapping lay or social-media-style statements to scholarly abstracts and by explicitly separating retrieval from three-way verification (Abu Ahmad et al., 2025a, 2025b). Its current distribution also includes narrative metadata derived from the CARDS hierarchy. CARDS organizes contrarian climate claims into broad categories and more specific subclaims, enabling analysis of recurring argumentative themes rather than a single flat misinformation label (Coan et al., 2021). Hierarchical climate-misinformation research likewise finds value in retaining taxonomy

structure when identifying narrative stimuli (Rojas et al., 2024). In the present design, these codes define analysis groups and a binary auxiliary target; they do not replace the official verification labels.

Retrieval quality is consequential because a verifier cannot reason from evidence that has not been surfaced. BM25 remains a strong sparse baseline due to its term-frequency saturation and document-length normalization (Robertson & Zaragoza, 2009). Dense passage retrieval and Siamese sentence encoders address vocabulary mismatch by mapping queries and passages into a shared semantic space (Karpukhin et al., 2020; Reimers & Gurevych, 2019). Recent fact-verification work further argues that evidence retrieval is often the dominant bottleneck, supporting careful measurement of ranking before downstream label prediction (L. Zheng et al., 2024).

ClimateCheck systems exemplify the multi-stage response to this bottleneck. The winning entry combined BM25, BGE reranker ensembles, hard-negative learning, rank fusion, and LLM-based analysis (J. Wang et al., 2025). Another system combined lexical, sparse-neural, dense-neural, cross-encoder, and generative stages (Kiepura & Lam, 2025). These architectures demonstrate the effectiveness of specialization across stages, but they also complicate energy accounting because each candidate may pass through several costly components. A resource-sensitive comparison of LLMs and BERT-based classifiers for social-media claim verification found that model-family choices entail material efficiency trade-offs (Upravitelev et al., 2025). SmolLM2 was developed as a data-centric small-language-model family for resource-constrained deployment (Ben Allal et al., 2025). The present experiments include a quantized 135M-parameter member as a local zero-shot ranking comparator; larger multi-stage ClimateCheck systems remain prior-work reference points because their retrieval pools and hardware boundaries differ.

For verification, fluent generation is not itself evidence of correctness. Candidate selection, relation classification, citation quality, and uncertainty should remain separately observable. This motivates a compact pipeline whose representations and probabilities can be inspected directly. BM25 and TF–IDF provide lexical views; truncated singular-value decomposition supplies a latent-semantic view; pair features encode overlap, quantities, negation, and length; and the narrative detector contributes a predicted contextual signal. A class-conditional character 3–5-gram model adds a probabilistic paired-text baseline with minimal inference cost.

Probability quality is as important as label accuracy when predictions govern escalation. Modern classifiers can be overconfident, and temperature scaling offers a simple post-hoc correction fitted on held-out logits (Guo et al., 2017). Expected calibration error is useful but sensitive to binning, which supports equal-count bins, reporting bin populations, and interval estimates when high-confidence bins contain few observations (Roelofs et al., 2022). The Brier score complements calibration-gap summaries by evaluating the full probability vector as a proper scoring rule (Brier, 1950).

Selective classification converts calibrated confidence into an operational decision: predict on retained cases and abstain on the remainder. Risk–coverage curves reveal whether lower coverage genuinely concentrates errors among deferred instances (Geifman & El-Yaniv, 2017). In scientific fact checking, abstention is not a failure state; it can route ambiguous claim–abstract relations to specialists or trigger a search for additional evidence. The value of this mechanism depends on threshold discipline. A threshold chosen on test labels overstates achievable coverage, so the present study locks candidate thresholds using only grouped out-of-fold training predictions.

Efficiency completes the evaluation. Energy use depends on hardware, runtime, memory, and the measurement boundary, and transparent reporting should accompany model-quality comparisons (Henderson et al., 2020). Green-computing methods recommend normalizing consumption by a clearly defined workload while documenting whether power was directly measured or estimated (Lannelongue et al., 2021). The experiments therefore report fitting separately from warm inference and normalize the latter to 1,000 claim–abstract pairs. This protocol does not assume that the lowest-energy model is preferable in every setting; rather, it exposes whether a quality gain is accompanied by a small, moderate, or disproportionate increase in computation.

Taken together, recent evidence-centered systems identify several distinct reliability layers: candidate discovery, ordering, relation assessment, provenance, and stopping. Rank-aggregation theory supplies a complementary perspective in which the stability of an ordering depends on how noisy preference signals are combined (Zhong & Ling, 2024a). This distinction is important for ClimateCheck because a high relation score cannot repair a missing abstract, and a good rank does not by itself establish that an answer is sufficiently supported. The present dataset directly measures candidate ordering and three-way relation prediction; calibrated abstention is a limited stopping mechanism, whereas multi-document provenance remains future work.

Cross-domain retrieval systems illustrate how evidence boundaries change with the operational setting. Microservice and cloud copilots ground root-cause summaries in fault signals, alerts, or private runbooks (G. Liu et al., 2023; G. Liu et al., 2024; B. Zhang et al., 2024). Code retrieval and normal-log prototypes likewise separate finding a relevant artifact from explaining it (Y. Li, 2024; Xin, 2025a). More recent HDFS, financial-search, and AIOps studies extend the same separation through deterministic failure narratives, hybrid reranking, and cost-aware routing (C. Li et al., 2026; Meng et al., 2026; Sun et al., 2026). These applications do not establish climate-domain accuracy, but they converge on a useful principle: evidence selection, explanatory text, and operational action should remain separately auditable.

The separation is especially visible when a retrieved answer can trigger a costly action. DevOps and quantitative assistants add command-safety or numerical guardrails after retrieval and generation (Z. Li et al., 2026; B. Zhang et al., 2026). Accounting-aware retrieval, evidence-constrained monitoring, and multi-regulation legal RAG restrict conclusions to defined documentary and jurisdictional scopes (Y. Chen et al., 2025; J. Li & Zhou, 2026; S. Zhou et al., 2026). Review-grounded recommendation and therapist-facing retrieval similarly use observations or dialogue history to justify rather than merely generate a response (Chang et al., 2026; Y. Zhang & H. Zhang, 2025a). For climate verification, the transferable lesson is provenance discipline, not the transfer of application-specific performance.

Evidential correctness is also distinct from adversarial safety. Continual red-teaming, evidence-based self-verification, and behavior-level refusal study how guardrails adapt to jailbreak or toxic prompts while preserving legitimate utility (D. Zheng & Li, 2024; D. Zheng et al., 2023; D. Zheng et al., 2024). Privacy and data-integrity risk cards extend this concern to human oversight of prompt-injected agents (Su, Rao, et al., 2026), while trajectory reliability and attack-chain modeling expose process failures that a final-answer score can miss (Bai et al., 2026; Zhong et al., 2026). Work on dark-pattern explanation further shows that apparently benign language can encode manipulative behavior requiring an explicit detection layer (Xu et al., 2025). These precedents motivate audit trails without turning ClimateCheck into a cybersecurity benchmark.

Security analytics also demonstrates the value of localized, traceable evidence. Language-guided feature selection, system-call window localization, and deep autoencoding associate a decision with observable traffic or execution features (Y. Li & Lu, 2025; Xin, 2026c; Xin et al., 2024). Predictive DDoS mitigation, LogBERT-style detection, and conformal alert control add operational response, self-supervised representation learning, or calibrated escalation to that trace (B. Wang et al., 2024; Xin, 2026d, 2026f). The analogy to claim verification is methodological: a system should reveal which evidence changed the decision and when confidence is insufficient for autonomous action.

Calibration becomes operational when an error changes what receives review. Credit, fraud, and bankruptcy studies combine probability calibration, uncertainty-driven rejection, cost sensitivity, and class-aware explanations, showing that aggregate accuracy can conceal consequential minority-class failures (Y. Chen et al., 2023; Y. Chen et al., 2024; Jin et al., 2024a, 2024b). Conformal GPU demand envelopes, residual fusion, and calibrated sensor fusion address different sources of uncertainty but share the need to connect confidence with a bounded decision (S. Chen et al., 2024; Nie & Zheng, 2024; Xin, 2025c). Medical vision–language classification, counterfactual credit scoring, and group-synchronization theory further distinguish model, subgroup, and latent-structure uncertainty (Lu & Zou, 2026; Xin, 2026b; Zhong & Ling, 2024b). These precedents support reporting calibration and selective risk beside macro-F1; their domain-specific guarantees are not imported into this study.

Distribution shift changes both representation quality and confidence. Cross-cloud transfer and few-shot cold-start forecasting examine workload and topology changes, while approximately shared-feature theory formalizes when information can bridge domains (He et al., 2024; He et al., 2025; Zhong, Pan, et al., 2025). Subject-independent BCI and cross-dataset activity recognition demonstrate related transfer difficulties across people and sensor collections (Wu et al., 2025; J. Zhang, 2025). News-based macro-market fusion supplies a temporal-regime example (H. Zhou & K. Zhang, 2025), and self-supervised customer representations show how compact features can be reused across downstream tasks (Xin, 2026e). The present study does not conduct external-domain transfer; grouped cross-fitting, novel-abstract sensitivity, and narrative slices probe narrower dependence within ClimateCheck.

A calibrated score becomes useful to a reviewer only when its evidential basis is visible. Scientific-search evidence cards, accounting-disclosure review cards, and financial dashboards organize source hierarchy, chart context, and provenance for inspection (Jin, 2025b; Z. Li et al., 2025; K. Zhang et al., 2025). Medical explanation and safety-card studies add uncertainty, subgroup risk, and escalation cues for high-stakes interfaces (C. Li et al., 2025; Ye et al., 2025; Zhong, Wu, et al., 2025; B. Zhou, Li, et al., 2025). These works motivate a review destination for abstained climate claims, but the current experiment evaluates no interface or user outcome.

Other interface studies reinforce that explanation quality depends on information structure rather than fluent wording alone. Recommendation and infrastructure cards translate confidence and supporting features into editable or reviewable units (G. Liu et al., 2025; B. Zhang et al., 2025), while design-critic and counseling-card work makes human judgment and evidence-linked handoff explicit (Y. Li et al., 2025; Y. Zhang & H. Zhang, 2025b). Trust-calibrated multilingual RAG adds an access dimension by pairing retrieval with uncertainty communication across languages (Y. Chen & Xu, 2026). A future ClimateCheck interface could adopt these design principles only after separate usability and outcome evaluation.

Human oversight also depends on what contextual information a system stores and reuses. Confidence-gated dialogue memory treats writing, updating, and forgetting as controlled operations, while federated preference learning combines knowledge grounding with privacy constraints (Kuo et al., 2025; Sun et al., 2023). Rubric-grounded essay scoring, privacy-safe aggregate marketing models, and strategy-controlled emotional support each pair a prediction with explicit criteria or governance boundaries (Bai & Wu, 2026; Xin, 2026a; Y. Zhang & Z. Zhou, 2026). For climate verification, these are cross-domain architectural lessons: narrative context should remain auxiliary and auditable, and it should never substitute for the claim–evidence relation.

3. Methodology

3.1 Data and version control.

The analysis used the ClimateCheck train and test pair files and applied data_version == "2025" before fitting any representation or model. The filter retained 1,144 training pairs and 1,904 test pairs, exactly reproducing the 3,048 pairs described for ClimateCheck 2025 (Abu Ahmad et al., 2025a). The training portion contained 259 exact claim strings, 252 distinct claim identifiers, and 989 distinct abstracts; the test portion contained 176 exact claim strings, 172 identifiers, and 1,685 abstracts. Training labels comprised 446 Supports, 241 Refutes, and 457 Not Enough Information instances. Test labels comprised 749 Supports, 266 Refutes, and 889 Not Enough Information instances.

The audit found no null fields, no repeated (claim_id, abstract_id) pairs, no repeated exact claim–abstract pairs, and no train–test overlap in exact claim strings or claim identifiers. It also identified 1,879 training rows assigned to the 2026 version, which were excluded before vocabulary construction. Two hundred abstract identifiers occurred in both retained splits, affecting 284 test rows. This overlap does not expose the test claim labels, but it motivated a novel-abstract analysis on the remaining 1,620 test rows. Five training identifiers and four test identifiers corresponded to more than one exact claim string. Exact claim text was therefore the primary semantic query unit, while claim_id was the grouping unit for cross-fitting and resampling so that identifier-linked variants stayed together.

Table 1. Dataset composition and the effect of the strict 2025 version filter.

Split	Quantity	Value
train	rows	1,144
train	exact_claims	259
train	claim_ids	252
train	abstract_ids	989
train	supports	446
train	refutes	241
train	not_enough_information	457
train	candidate_min	1
train	candidate_mean	4.417
train	candidate_median	5
train	candidate_max	5
train	claims_with_evidence	230
train	claims_0_0	176
train	claims_any_narrative	83
test	rows	1,904
test	exact_claims	176
test	claim_ids	172
test	abstract_ids	1,685
test	supports	749
test	refutes	266
test	not_enough_information	889
test	candidate_min	1
test	candidate_mean	10.82
test	candidate_median	11
test	candidate_max	17
test	claims_with_evidence	166
test	claims_0_0	127
test	claims_any_narrative	49

Table 2. Identifier integrity, candidate-pool properties, and train–test overlap.

Check	Value
unfiltered_train_rows	3,023
included_train_2025_rows	1,144
excluded_non_2025_train_rows	1,879
test_rows_all_2025	1,904
train_null_cells	0
test_null_cells	0
train_duplicate_claim_id_abstract_id	0
test_duplicate_claim_id_abstract_id	0
train_duplicate_exact_claim_abstract	0
test_duplicate_exact_claim_abstract	0
exact_claim_overlap	0
claim_id_overlap	0
shared_abstract_ids	200
test_rows_with_shared_abstract	284
train_claim_ids_with_text_collisions	5
test_claim_ids_with_text_collisions	4

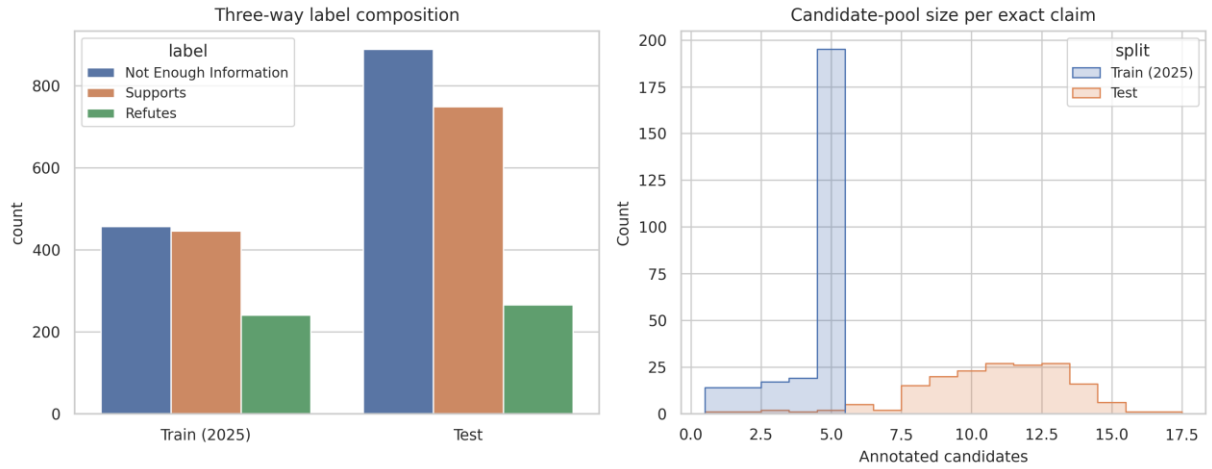

Figure 1. Label composition and candidate-pool size in the 2025 train and test splits.

Task definitions. Candidate ranking was defined within each exact claim's annotated test pool. For claim q, the candidate set D_q consisted only of abstracts paired with that exact text. A candidate was relevant when its label was Supports or Refutes and nonrelevant when its label was Not Enough Information:

$$r(q,d) = \mathbb{1}\{y(q,d) \in [\text{Supports, Refutes}]\}.$$

The primary ranking population contained 166 exact test claims with at least one evidentiary candidate, represented by 163 claim-identifier blocks. Ten all-Not-Enough-Information claims were omitted from MAP and the principal NDCG@5 calculation

because average precision is undefined without relevant items. An all-query NDCG@5 analysis assigned zero to these ten claims. Ties were resolved by ascending abstract_id. Mean average precision followed the standard query-level definition (Manning et al., 2008); NDCG@5 used binary gains and logarithmic discounting (Järvelin & Kekäläinen, 2002). Recall@2, Recall@5, Bpref, and mean reciprocal rank supplemented the two primary ranking metrics.

Verification treated each of the 1,904 test pairs as a three-class observation. The fixed class order was Not Enough Information, Refutes, and Supports. Macro-F1 was primary because it gives each class equal influence despite unequal support (Sokolova & Lapalme, 2009). Weighted F1, accuracy, balanced accuracy, multiclass Matthews correlation, negative log likelihood, multiclass Brier score, adaptive expected calibration error, and maximum calibration error were also computed. Classwise precision, recall, and F1 made errors on the smaller Refutes class visible.

3.2 Representations and baselines.

Text was normalized with Unicode NFKC, lowercased, and tokenized with the expression (?u)\b\w\w+\b; negations were retained and stemming was not applied. Every learned vocabulary and decomposition was fitted on training text only. The empirical suite included the following systems.

Seeded Random generated 100 deterministic candidate permutations per query with seeds 2025–2124 and served as a ranking floor. Its classification counterpart predicted the most frequent training label, while a fixed empirical-prior probability vector supported probability-score comparisons. BM25 used k_1=1.2 and b=0.75, with document frequencies estimated from the 989 unique training abstracts (Robertson & Zaragoza, 2009). Query terms absent from this training index contributed zero.

Word TF–IDF Cosine used a shared claim-and-abstract vectorizer with word unigrams and bigrams, min_df=2, max_df=0.98, a 30,000-feature cap, sublinear term frequency, Unicode accent stripping, and L2 normalization. LSA-128 Cosine applied truncated singular-value decomposition to that matrix, using at most 128 components, seven iterations, and seed 2025. Dense LSA vectors were L2-normalized before cosine scoring. Although latent semantic representations share the independent-encoding efficiency goal of modern bi-encoders (Karpukhin et al., 2020; Reimers & Gurevych, 2019), LSA-128 is a linear latent-semantic encoder and not a transformer or LLM.

The causal-LM comparator used the Apache-2.0 SmolLM2-135M multilingual base weights in Q4_K_M GGUF form (Ben Allal et al., 2025). A zero-shot prompt asked which candidate gave the strongest evidence for or against the claim. To fit each complete pool inside 128 tokens, the claim retained its first six content words and each abstract contributed the first content word shared with the claim, or otherwise its first content word. Constrained decoding selected one index at temperature 0 and seed 2025; that candidate was promoted to rank one and BM25 ordered the remainder. One sequence and one CPU thread were used. This ranking-only stress test imposed a deliberately severe information boundary and produced no three-way verification probabilities.

The Compact N-gram LM used CountVectorizer character 3–5-grams and MultinomialNB with additive smoothing $\alpha=1.0$. Its input was the exact paired representation claim [PAIR] abstract. The model estimated a class-conditional n-gram distribution for each verification label and selected the class with the greatest posterior score. For ranking, its evidentiary score was the summed posterior probability of Supports and Refutes. Word+Char TF–IDF Logistic concatenated sparse word features with character char_wb features and fitted a class-balanced logistic classifier.

The Lexical–Semantic Logistic model used scalar pair features: BM25; word and character TF–IDF cosine; LSA cosine; token Jaccard overlap; claim-token coverage; shared-number count; a number-mismatch flag; negation indicators and their exclusive-or mismatch; claim and abstract lengths; log length ratio; and interactions between cosine, negation mismatch, and number agreement. Character TF–IDF used three- to five-character boundary-aware n-grams, min_df=2, a 20,000-feature cap, and sublinear term frequency. Features were standardized using training means and standard deviations. The binary reranker used class-balanced liblinear logistic regression with C=1.0, a 2,000-iteration limit, and seed 2025. The three-way verifier used class-balanced multinomial logistic regression with C=1.0, the L-BFGS solver, the same iteration limit, and the same seed. These models operationalize the retrieval-centered view that fine-grained evidence matching is decisive for verification (L. Zheng et al., 2024).

Table 3. Model definitions, selected hyperparameters, and fitted representation sizes.

Model	Representation	Fitted Feature Count	Learned Parameter Count	Fixed Hyperparameters	Artifact Size MB
Seeded Random	100 deterministic within-query permutations	0	0	seeds 2025–2124	0.000
BM25	training-abstract document frequencies	15,291	15,291	k1=1.2; b=0.75	0.101
Word TF-IDF Cosine	shared claim/abstract word 1–2-gram TF-IDF	30,000	30,000	min_df=2; max_df=0.98; max_features=30000; sublinear_tf	0.324
LSA-128 Cosine	Truncated SVD over shared word TF-IDF	128	3,840,000	n_components=128; n_iter=7; seed=2025	27.96
Compact N-gram LM	class-conditional paired-text character 3–5-gram counts	30,000	90,003	MultinomialNB additive alpha=1.0; max_features=30000	0.682
Word+Char TF-IDF Logistic	paired-text word 1–2-gram + char 3–5-gram TF-IDF	50,000	150,003	C=1; balanced; lbfgs/liblinear; OOF temperature/Platt	1.810
Lexical-Semantic Logistic	17 standardized lexical/semantic scalar features	17	54	C=1; balanced; grouped 5-fold OOF calibration	28.24
Narrative-Aware Logistic	lexical-semantic features + cross-fitted predicted narrative interactions	24	75	claim-only binary detector; C=1; predicted-group temperatures	28.32
SmolLM2-135M top-choice + BM25	135M-parameter Q4_K_M causal LM; constrained zero-shot top-candidate choice; one claim-conditioned content word per abstract; BM25 orders the remainder	135,000,000	135,000,000	128-token context; temperature=0; seed=2025; 1 lane; 1 CPU thread	100.6

3.3 Narrative-aware modeling.

Narrative strings were treated as multi-label claim metadata. The auxiliary target was binary: z(q)=1 when a claim carried any nonzero narrative code and z(q)=0 for 0_0. A claim-only detector concatenated word and character TF–IDF features and fitted class-balanced logistic regression. Five-fold StratifiedGroupKFold with shuffling and seed 2025 produced an out-of-fold probability p(z=1| q) for every training claim; groups were claim_id, preventing linked claim variants from crossing folds. The detector was then refitted on all training claims to predict test narrative probabilities.

The Narrative-Aware Lexical–Semantic Logistic model added this probability and its products with BM25, word cosine, character cosine, LSA cosine, negation mismatch, and number agreement. Gold test narrative codes were never used as predictive inputs. A guard test replaced the test narrative field with a constant and confirmed identical predictions. Gold codes were used only after prediction to describe results for 0_0, any nonzero narrative, and top-level categories 1–5. A claim with multiple semicolon-delimited codes contributed to each applicable subgroup rather than being forced into its first category. This preserves the hierarchical, overlapping character of the taxonomy (Coan et al., 2021; Rojas et al., 2024).

3.4 Cross-fitting and calibration.

All hyperparameters were declared before test evaluation. Five claim-identifier-grouped outer folds generated out-of-fold logits for calibration and threshold selection. Within every outer fold, count and TF–IDF vocabularies, document frequencies, the SVD

decomposition, scalar-feature standardizers, and classifiers were fitted only on that fold's training partition before transforming its held-out claims. Narrative-aware outer-fold features used an additional claim-grouped inner cross-fit: each outer-training claim received a detector probability from an inner model that had not seen its identifier, and the detector applied to the outer validation fold was fitted only on outer-training claims. After the strict fold-local procedure, each uncalibrated representation and estimator was refitted on all 1,144 training pairs and evaluated once on the untouched test set. No test label influenced a vocabulary, decomposition, scaler, coefficient, temperature, or confidence threshold.

Three probability variants were compared: uncalibrated softmax, one global temperature, and predicted-narrative-conditioned temperatures. The global temperature $T \in [0.05,10]$ minimized out-of-fold negative log likelihood under bounded scalar optimization (Guo et al., 2017). Narrative-conditioned calibration split out-of-fold pairs by the predicted narrative class at probability 0.5 and fitted a temperature for a group only when it contained at least 100 pairs and all three verification labels; otherwise it inherited the global temperature. The same predicted, rather than gold, grouping rule was applied at test time. Binary relevance logits were calibrated with Platt scaling.

Multiclass Brier score was the mean squared distance between each probability vector and its one-hot label (Brier, 1950). Adaptive ECE used ten equal-count bins and weighted the absolute confidence–accuracy gap by bin size. Reliability plots encoded bin counts so that sparse high-confidence regions remained visible. Equal-count binning and interval reporting reduce, though do not eliminate, finite-sample bias in calibration estimates (Roelofs et al., 2022).

3.5 Selective abstention.

Confidence was the largest calibrated class probability. For a threshold τ, coverage was the retained fraction and selective risk was one minus accuracy among retained pairs. The complete curve and its area, AURC, were computed. Operating thresholds for target risks of 10% and 20% were screened exclusively on out-of-fold training predictions. Thresholds from 0 to 1 in increments of 0.001 were eligible when their 95% claim-cluster bootstrap upper risk bound met the target while retaining at least 30 claim groups and 100 pairs. The lowest eligible threshold was locked before test evaluation; otherwise the policy was recorded as infeasible. Because the same OOF predictions supported calibration and a 1,001-threshold scan, this is an empirical bootstrap screen rather than a simultaneous risk guarantee. Retained-set macro-F1 accompanied risk and coverage. The protocol follows the selective-classification principle that performance should be exposed as a function of coverage rather than forced on every case (Geifman & El-Yaniv, 2017).

3.6 Uncertainty and hypothesis tests.

Metric intervals used 1,000 paired cluster-bootstrap replicates, sampling test claim_id groups with replacement and retaining every pair in each sampled group. The ranking comparison averaged exact-query AP differences within claim-identifier blocks before 10,000 two-sided sign flips. Macro-F1, ECE, and AURC differences were calculated within the same paired claim-block bootstrap replicates, and two-sided bootstrap tail probabilities were reported. This paired design follows evidence that query-level randomization and bootstrap procedures are appropriate for information-retrieval comparisons (Smucker et al., 2007). Four confirmatory contrasts were specified: narrative-aware reranking versus BM25 for MAP; narrative-aware verification versus Word+Char TF–IDF Logistic for macro-F1; narrative-conditioned versus global temperature for ECE; and the same calibration comparison for AURC. Holm adjustment controlled familywise error. A separate exploratory accuracy comparison used 10,000 claim-identifier-block sign flips.

3.7 Efficiency and reproducibility.

Warm inference began with raw text and included narrative prediction, vectorization, ranking, verification, calibration, and output. Package import and model loading were excluded, while one-time fitting cost was reported separately. Runs used one process and one CPU thread, with OpenMP, MKL, OpenBLAS, and NumExpr thread counts fixed at one. For the fitted local systems, two warm-up passes preceded five complete-test inference blocks executed in seeded randomized order; each block processed one or more full 1,904-pair passes for at least 0.1 seconds. SmolLM2 inference was timed for each complete query pool with the same one-thread boundary, and the 176 query-batch measurements were normalized by their candidate-pair counts. The logs recorded the exact number of scored pairs, wall time, and process CPU time. A standardized energy proxy was calculated from CPU time at a constant 15 W:

$$E_{proxy,1000} = (15 \times t_{CPU} \times 1{,}000) / (3.6 \times 10^6 \times n) \text{ kWh}.$$

This value is a workload-normalized CPU-time proxy, not an electricity-meter reading. CPU time and wall time are therefore reported beside it. Ranking-only inference paths and paths producing both ranking and verification outputs were identified separately so that unequal workloads were not treated as interchangeable. The table's fit time records the final component fit rather than the complete grouped cross-fitting procedure. The hardware boundary was an AMD EPYC 9V74 80-Core Processor

with nine available logical cores and 21.9 GB RAM; numerical thread variables were fixed to one. Artifact size, software versions, and median and interquartile range across blocks were retained, consistent with systematic energy-reporting guidance (Henderson et al., 2020; Lannelongue et al., 2021). Prediction files stored row identifiers, raw logits, calibrated probabilities, confidence, ranking scores, predicted narrative probabilities, fold assignments, and bootstrap seeds.

4. Results and Discussion

4.1 Data composition and audit.

Filtering on data_version == "2025" retained 1,144 training pairs and excluded 1,879 rows from the later extension. The held-out set contained 1,904 pairs: 889 Not Enough Information, 749 Supports, and 266 Refutes. Candidate pools were larger at test time, averaging 10.818 abstracts per exact claim versus 4.417 in training. Of 176 test claims, 166 had at least one evidentiary abstract and ten were entirely Not Enough Information. Evidence-bearing candidates represented 56.0% of pairs within the primary ranking queries, which explains the relatively high expected MAP of an unordered ranking. The absence of train–test claim overlap supported a clean claim-level evaluation, while the 200 shared abstract identifiers motivated a separate novel-abstract analysis.

4.2 Closed-pool evidence ranking.

Table 4 reports claim-cluster intervals, and Figure 2 compares MAP with NDCG@5. The seeded random control obtained MAP 0.646, BM25 0.649, Word TF–IDF Cosine 0.646, and LSA-128 Cosine 0.665. SmolLM2 top-choice plus BM25 reached MAP 0.647 (95% CI [0.607, 0.685]) and NDCG@5 0.599 (95% CI [0.553, 0.643]). Its Recall@2, Recall@5, Bpref, and MRR were 0.210, 0.494, 0.573, and 0.721. The character n-gram classifier reached MAP 0.655, Lexical–Semantic Logistic 0.653, and Narrative-Aware Logistic 0.657 (95% CI [0.619, 0.693]). Word+Char TF–IDF Logistic was the strongest reranker, with MAP 0.670 and NDCG@5 0.639. The narrative-aware model produced NDCG@5 0.619, Recall@2 0.215, Recall@5 0.513, Bpref 0.581, and MRR 0.700.

Table 4. Closed-pool evidence-ranking performance on the test split.

Model	MAP [95% Ci]	NDCG At 5 [95% Ci]	Recall2	Recall5	Bpref	MRR
Seeded Random	0.646 [0.616, 0.676]	0.600 [0.569, 0.632]	0.205	0.483	0.568	0.713
BM25	0.649 [0.613, 0.686]	0.595 [0.554, 0.640]	0.212	0.491	0.577	0.686
Word TF-IDF Cosine	0.646 [0.606, 0.682]	0.602 [0.557, 0.647]	0.198	0.510	0.570	0.673
LSA-128 Cosine	0.665 [0.624, 0.706]	0.620 [0.573, 0.665]	0.222	0.497	0.602	0.713
SmolLM2-135M top-choice + BM25	0.647 [0.607, 0.685]	0.599 [0.553, 0.643]	0.210	0.494	0.573	0.721
Compact N-gram LM	0.655 [0.616, 0.691]	0.616 [0.573, 0.658]	0.201	0.505	0.578	0.705
Word+Char TF-IDF Logistic	0.670 [0.632, 0.707]	0.639 [0.596, 0.682]	0.226	0.524	0.585	0.725
Lexical-Semantic Logistic	0.653 [0.611, 0.688]	0.610 [0.563, 0.654]	0.209	0.497	0.578	0.714
Narrative-Aware Logistic	0.657 [0.619, 0.693]	0.619 [0.577, 0.661]	0.215	0.513	0.581	0.700

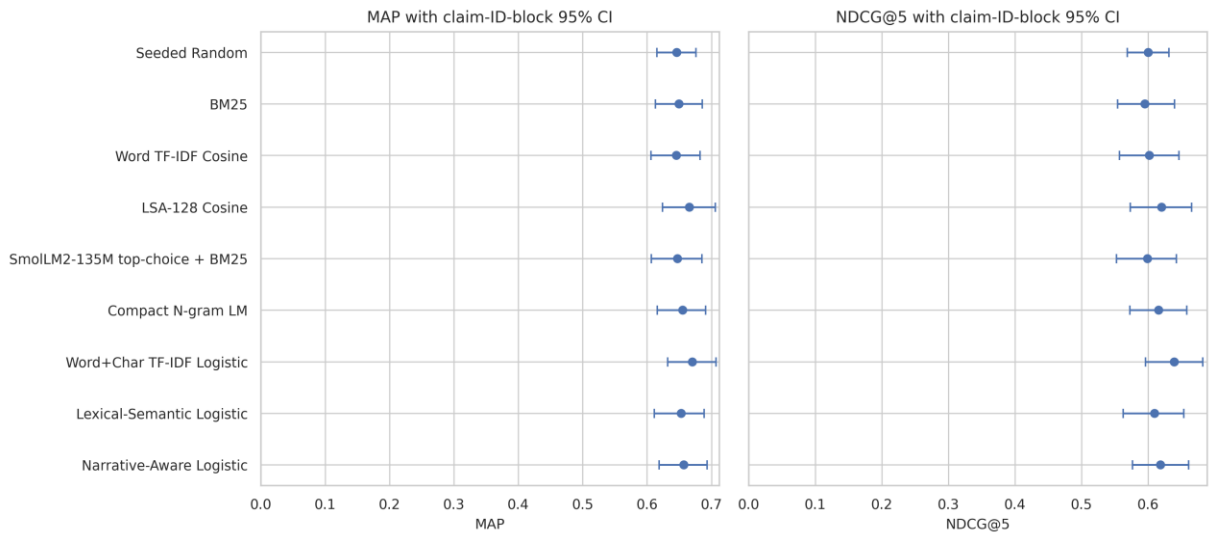

Figure 2. Closed-pool MAP and NDCG with claim-cluster bootstrap intervals.

Relative to BM25, narrative-aware MAP increased by 0.007, but the 95% paired interval [−0.023, 0.036] crossed zero and the Holm-adjusted p-value was 1.000. Narrative-aware scoring improved average precision on 75 queries, decreased it on 70, and tied on 21. The result is therefore a small descriptive gain over BM25, not evidence of uniform superiority; the sparse Word+Char model remained stronger overall. SmolLM2's MRR exceeded BM25's 0.686, indicating more first-relevant hits, but its MAP was 0.002 lower because the causal model altered only the top position and BM25 ordered every remaining candidate. Given the one-word abstract representation, this row is an information-boundary stress test rather than a full-text estimate of small-LM capability. These values describe ordering inside annotated candidate pools and are not comparable to the open-corpus ClimateCheck leaderboard, which includes candidate generation from hundreds of thousands of publications (Abu Ahmad et al., 2025b).

The ranking pattern is consistent with the broader observation that elaborate retrieval pipelines do not dominate under every pool and cost boundary. Hybrid reranking can improve the ordering of financial or operational evidence, and retrieval-quality models can flag weak candidate sets before generation (Meng et al., 2026; R. Zhang et al., 2025). Here, however, the sparse Word+Char representation led within a small annotated pool. That result does not argue against multi-hop retrieval or evidence-chain modeling in an open corpus; it identifies the strongest measured option for the present candidate boundary.

4.3 Three-way verification.

Aggregate results appear in Table 5, classwise scores in Table 6, and normalized confusion matrices in Figure 7. The empirical-prior classifier obtained macro-F1 0.212. Macro-F1 was 0.387 for the character n-gram classifier, 0.392 for Word+Char TF–IDF Logistic, 0.371 for Lexical–Semantic Logistic, and 0.394 (95% CI [0.350, 0.435]) for Narrative-Aware Logistic. The narrative-aware model's accuracy was 0.404, balanced accuracy 0.472, weighted F1 0.402, and MCC 0.159. Word+Char achieved higher raw accuracy, 0.430, whereas the narrative-aware model distributed recall more evenly across classes.

Table 5. Three-way claim–abstract verification performance.

Model	Accuracy [95% Ci]	Balanced Accuracy	Macro-f1 [95% Ci]	Weighted F1	MCC	NLL	Brier	ECE	AURC
Empirical Prior	0.467 [0.426, 0.510]	0.333	0.212 [0.199, 0.225]	0.297	0.000	1.017	0.617	0.093	0.568
Compact N-gram LM	0.397 [0.361, 0.432]	0.428	0.387 [0.349, 0.424]	0.404	0.096	18.43	1.196	0.596	0.593
Word+Char TF-IDF Logistic	0.430 [0.398, 0.462]	0.400	0.392 [0.361, 0.423]	0.434	0.084	1.053	0.638	0.035	0.530
Lexical-Semantic Logistic	0.389 [0.348, 0.427]	0.415	0.371 [0.332, 0.409]	0.395	0.109	1.097	0.662	0.043	0.580
Narrative-Aware Logistic	0.404 [0.360, 0.448]	0.472	0.394 [0.350, 0.435]	0.402	0.159	1.072	0.647	0.025	0.551

Table 6. Class-specific verification precision, recall, and F1.

Model	Class	Precision	Recall	F1	Support
Empirical Prior	Not Enough Information	0.467	1	0.637	889
Empirical Prior	Refutes	0	0	0.000	266
Empirical Prior	Supports	0	0	0.000	749
Compact N-gram LM	Not Enough Information	0.534	0.327	0.406	889
Compact N-gram LM	Refutes	0.237	0.523	0.326	266
Compact N-gram LM	Supports	0.422	0.435	0.429	749
Word+Char TF-IDF Logistic	Not Enough Information	0.515	0.511	0.513	889
Word+Char TF-IDF Logistic	Refutes	0.224	0.316	0.262	266
Word+Char TF-IDF Logistic	Supports	0.432	0.374	0.401	749
Lexical-Semantic Logistic	Not Enough Information	0.530	0.256	0.346	889
Lexical-Semantic Logistic	Refutes	0.187	0.470	0.267	266
Lexical-Semantic Logistic	Supports	0.483	0.518	0.500	749
Narrative-Aware Logistic	Not Enough Information	0.539	0.233	0.325	889
Narrative-Aware Logistic	Refutes	0.228	0.669	0.341	266
Narrative-Aware Logistic	Supports	0.518	0.513	0.515	749

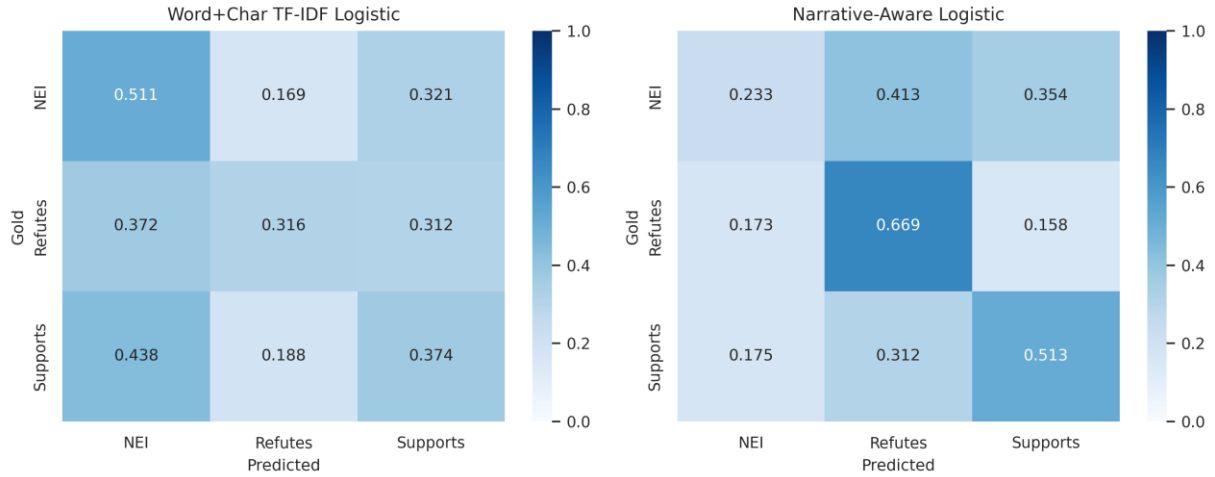

Figure 7. Row-normalized verification confusion matrices.

Narrative-aware F1 was 0.325 for Not Enough Information, 0.341 for Refutes, and 0.515 for Supports. Refutes recall reached 0.669 but precision was 0.228, while Not Enough Information recall was 0.233. The model consequently recovered many contradictions at the cost of assigning evidentiary labels to insufficient pairs. Relative to Word+Char, its macro-F1 difference was only +0.002 (95% CI [−0.037, 0.045], Holm-adjusted p=1.000). Thus, the higher balanced accuracy is a useful error-profile difference, not a statistically resolved improvement.

4.4 Calibration and abstention.

Table 7 and Figure 3 show that uncalibrated narrative-aware probabilities were overconfident: NLL was 1.118, Brier score 0.670, and adaptive ECE 0.105. Global temperature scaling improved these values to 1.083, 0.650, and 0.035. Predicted-narrative-conditioned scaling used temperatures 1.527 for the predicted 0_0 group and 5.223 for the predicted-narrative group, yielding NLL 1.072, Brier 0.647, and ECE 0.025. Its ECE difference from global scaling was −0.010 (95% CI [−0.030, 0.023], Holm-adjusted p=1.000). Reliability curves show the overall compression of overconfident probabilities and the residual bin-level deviations. Temperature scaling did not change class decisions or monotonic within-model ranking.

Table 7. Probability calibration before and after temperature scaling.

Calibration	NLL [95% Ci]	Brier [95% Ci]	ECE [95% Ci]	Mce	AURC [95% Ci]	Tempera ture 0 0	Tempera ture Narrativ e
Uncalibrated	1.118 [1.067, 1.175]	0.670 [0.642, 0.698]	0.105 [0.074, 0.149]	0.268	0.568 [0.500, 0.632]	1	1
Global temperature	1.083 [1.056, 1.113]	0.650 [0.636, 0.664]	0.035 [0.026, 0.074]	0.083	0.568 [0.499, 0.632]	2.141	2.141
Predicted-narrative temperature	1.072 [1.049, 1.094]	0.647 [0.633, 0.659]	0.025 [0.026, 0.068]	0.063	0.551 [0.482, 0.610]	1.527	5.223

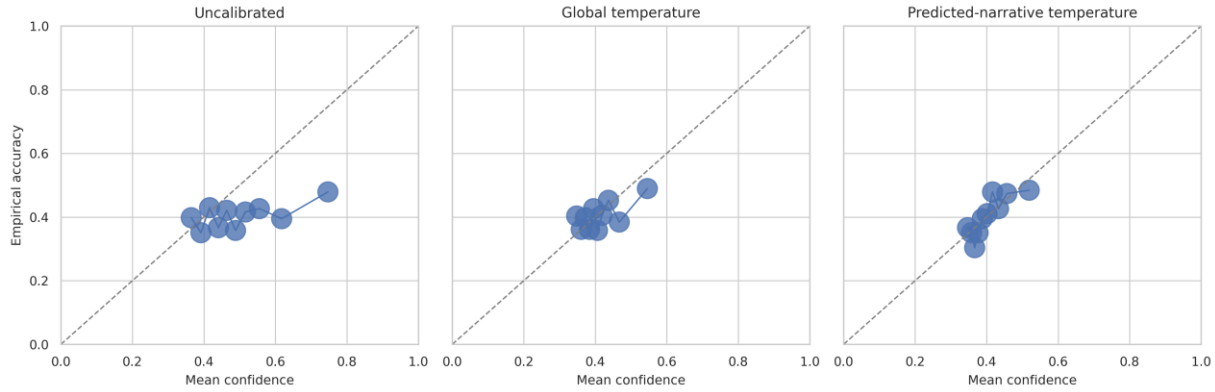

Figure 3. Reliability of verification probabilities before and after calibration.

The complete risk–coverage analysis is reported in Table 8 and Figure 4. Narrative-conditioned calibration achieved AURC 0.551, compared with 0.568 under both global scaling and the uncalibrated model. The difference from global scaling was −0.017 (95% CI [−0.064, 0.025], Holm-adjusted p=1.000). Neither the 10% nor the 20% target-risk rule produced a feasible operating point under the prespecified out-of-fold bootstrap screen. No test threshold was therefore applied for either target. This negative result is operationally informative: the observed confidence ordering improved the complete curve slightly, but it did not support the proposed low-risk automation policies.

The infeasible target-risk screens also argue against converting confidence into an automatic answer policy for this test set. Evidence and safety-card research suggests a safer use of uncertainty: expose the underlying source, probability, and escalation state instead of hiding deferred cases (Jin, 2025b; C. Li et al., 2025; Su, Rao, et al., 2026). A comparable ClimateCheck interface could present the highest-ranked abstract and a clear "additional evidence required" state, but that remains a design implication rather than an evaluated outcome.

Table 8. Selective prediction at fixed coverage and training-locked risk targets.

Target Risk	Status	Oof Locked Threshold	Test Coverage	Test Risk	Retained Pairs	Retained Macro F1	Full Test AURC
0.100	no feasible operating point	—	—	—	0	—	0.551
0.200	no feasible operating point	—	—	—	0	—	0.551

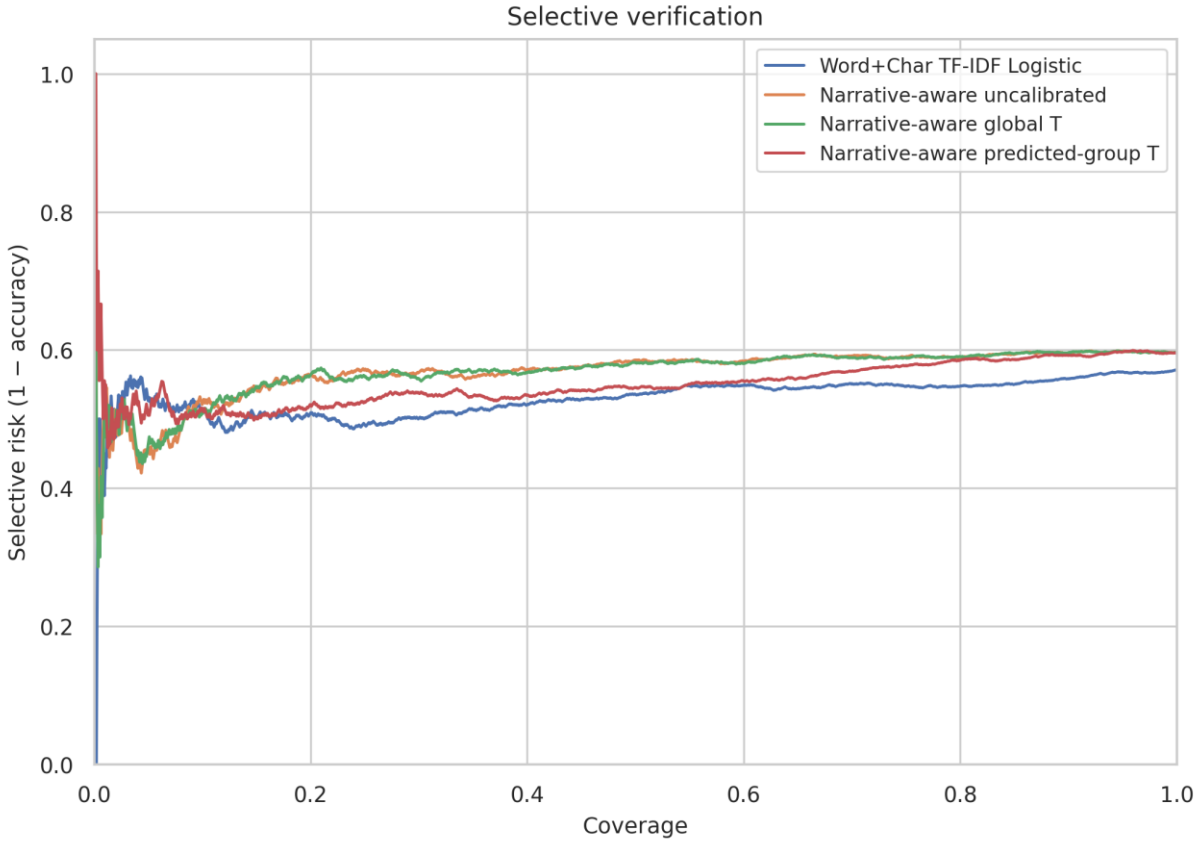

Figure 4. Selective risk as lower-confidence predictions are deferred.

4.5 Narrative subgroups.

Table 9 and Figure 5 separate 127 0_0 claims from 49 claims carrying at least one nonzero CARDS code. For 0_0, macro-F1 was 0.319, ECE 0.046, and MAP 0.656. For the nonzero-narrative group, the corresponding values were 0.396, 0.089, and 0.659. Across top-level categories, macro-F1 ranged from 0.291 to 0.473, ECE from 0.094 to 0.191, and MAP from 0.578 to 0.850. The category strata overlap for multi-label claims and several contain only six to sixteen exact claims; they locate heterogeneous error patterns but do not support causal claims about narrative membership.

Table 9. Performance across CARDS narrative strata.

Subgroup	Pairs	Exact Claims	Macro-f1 [95% Ci]	ECE [95% Ci]	MAP [95% Ci]
0	1,399	127.0	0.319 [0.283, 0.355]	0.046 [0.040, 0.091]	0.656 [0.613, 0.699]
Any nonzero narrative	505	49.00	0.396 [0.314, 0.462]	0.089 [0.068, 0.165]	0.659 [0.585, 0.730]
Top-level 1	176	16.00	0.473 [0.302, 0.582]	0.161 [0.106, 0.294]	0.625 [0.502, 0.737]
Top-level 2	167	16.00	0.291 [0.193, 0.379]	0.191 [0.145, 0.300]	0.677 [0.533, 0.812]
Top-level 3	53	6.000	0.397 [0.284, 0.477]	0.175 [0.157, 0.418]	0.850 [0.729, 0.947]
Top-level 4	64	6.000	0.420 [0.231, 0.511]	0.150 [0.119, 0.296]	0.589 [0.403, 0.763]
Top-level 5	112	11.00	0.361 [0.220, 0.469]	0.094 [0.093, 0.246]	0.578 [0.416, 0.735]

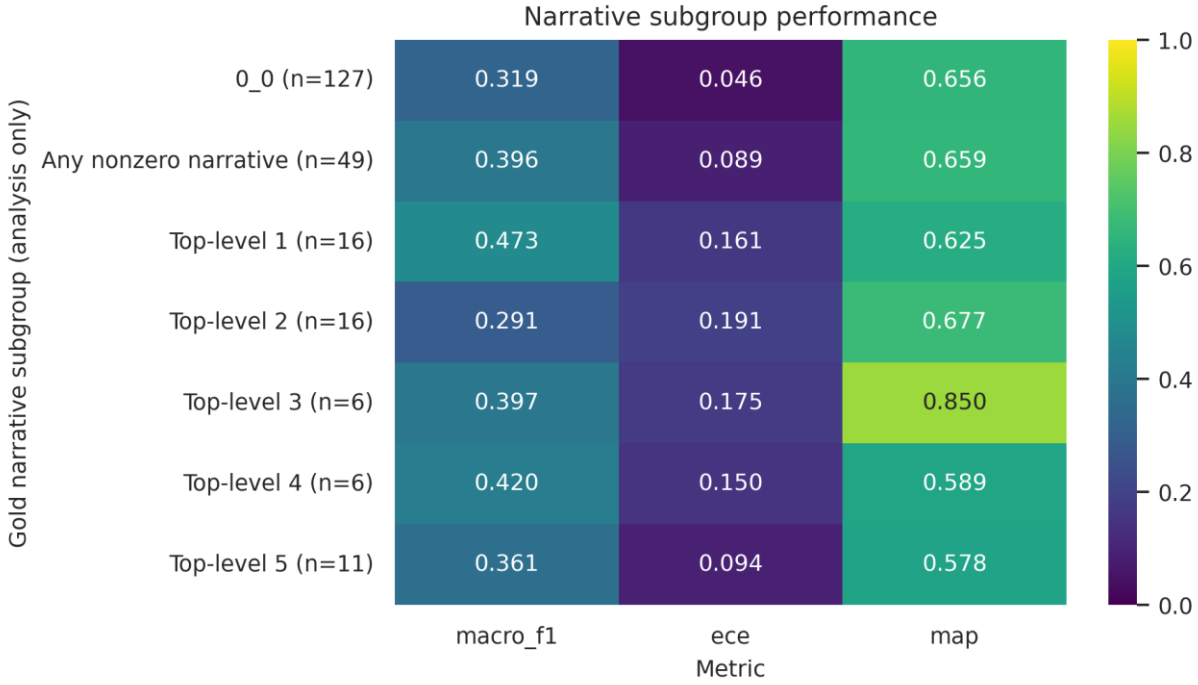

Figure 5. Narrative-stratified verification and ranking performance.

4.6 Ablation and robustness.

Table 10 shows metric-dependent feature contributions. Removing character similarity reduced MAP by 0.013 and macro-F1 by 0.005; removing LSA reduced them by 0.003 and 0.009. Removing negation and number features raised MAP by 0.002 but lowered macro-F1 by 0.008. The complete removal of narrative probabilities, interactions, and narrative-conditioned calibration changed MAP by −0.004, macro-F1 by −0.023, ECE by +0.017, and AURC by +0.029. Calibration itself accounted for an ECE reduction of 0.080 and an AURC reduction of 0.017. These ablations indicate that narrative information most clearly altered class balance and uncertainty ordering rather than maximizing ranking.

Table 10. Ablation of semantic, lexical, narrative, and calibration components.

Ablation	Retained Feature Count	MAP	Macro F1	ECE	AURC	Delta MAP Vs Full	Delta Macro F1 Vs Full	Delta ECE Vs Full	Delta AURC Vs Full
Full narrative-aware model	24	0.657	0.394	0.025	0.551	0.000	0.000	0.000	0
Remove character similarity	22	0.644	0.389	0.017	0.559	-0.013	-0.005	-0.008	0.008
Remove LSA similarity	22	0.654	0.385	0.028	0.553	-0.003	-0.009	0.002	0.003
Remove negation/number features	14	0.658	0.386	0.046	0.525	0.002	-0.008	0.020	-0.026
Remove narrative probability	23	0.658	0.399	0.023	0.547	0.001	0.005	-0.002	-0.004
Remove narrative interactions	18	0.655	0.386	0.027	0.550	-0.002	-0.008	0.002	-0.001

Ablation	Retained Feature Count	MAP	Macro F1	ECE	AURC	Delta MAP Vs Full	Delta Macro F1 Vs Full	Delta ECE Vs Full	Delta AURC Vs Full
Remove all narrative features	17	0.653	0.371	0.043	0.580	-0.004	-0.023	0.017	0.029
Remove calibration	24	0.657	0.394	0.105	0.568	0.000	0.000	0.080	0.017

Table 11 examines shared abstracts, query grouping, fold locality, and gold-narrative isolation. On the 1,620 pairs with novel abstract identifiers, MAP was 0.650, macro-F1 0.393, ECE 0.029, and AURC 0.558, closely tracking the complete test values. Pooling exact claim variants by identifier changed MAP from 0.657 to 0.647 over 163 evidence-bearing blocks. The predeclared seed 2025 produced the reported strict fold-local result, and replacing every gold test narrative value with a constant changed no prediction. Figure 8 displays the claim-block distributions for the four prespecified contrasts; each Holm-adjusted p-value was 1.000.

Table 11. Robustness checks and leakage-sensitivity analyses.

Analysis	Population	Pairs	MAP	Macro F1	ECE	AURC	Maximum Prediction Difference	Passed
Abstract-overlap sensitivity	Full test	1,904	0.657	0.394	0.025	0.551	—	—
Abstract-overlap sensitivity	Novel abstract IDs only	1,620	0.650	0.393	0.029	0.558	—	—
Ranking query-unit sensitivity	Exact claim text	1,904	0.657	—	—	—	—	—
Ranking query-unit sensitivity	claim_id pooled	1,904	0.647	—	—	—	—	—
Strict fold-local calibration	predeclared seed=2025	1,904	0.657	0.394	0.025	0.551	—	—
Gold-narrative invariance unit test	Test narrative replaced by constant	1,904	0.657	0.394	0.025	0.551	0	True

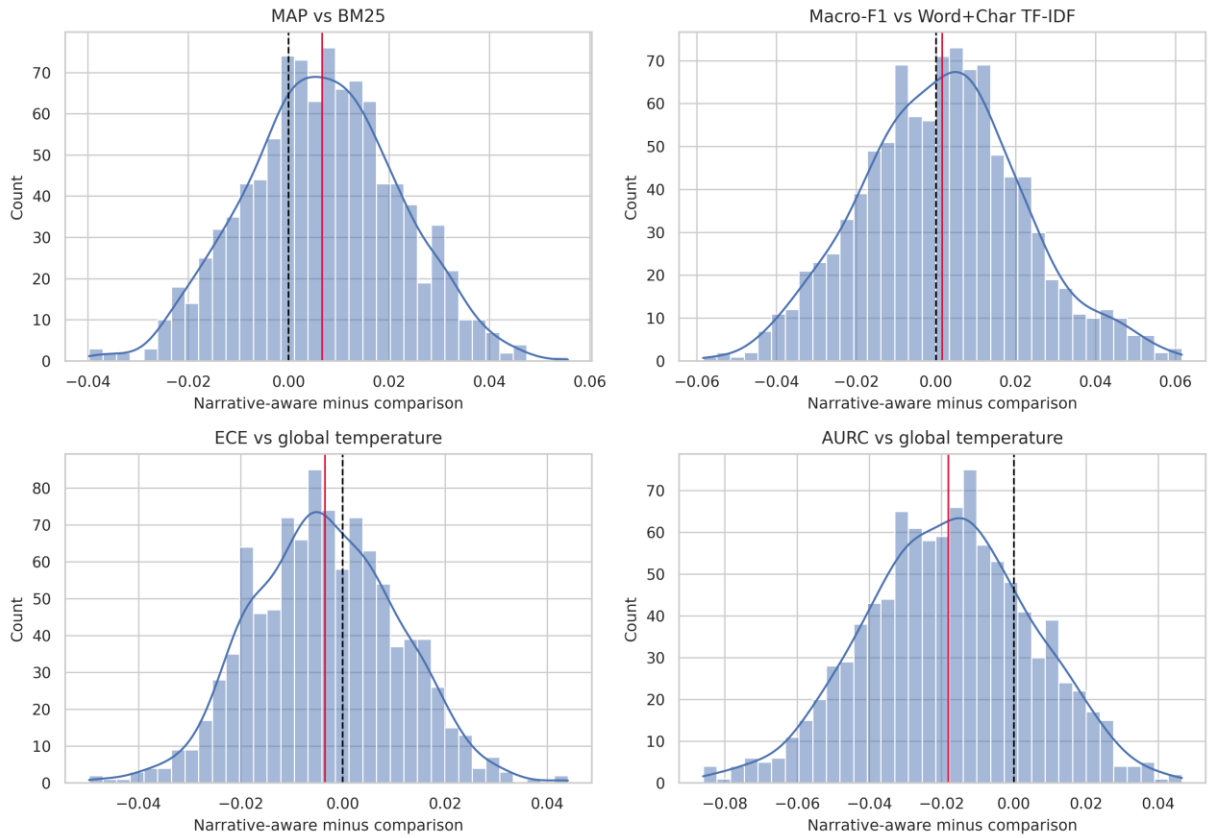

Figure 8. Claim-cluster bootstrap distributions of paired model improvements.

4.7 Efficiency and quality trade-offs.

Table 12 distinguishes ranking-only paths from pipelines that also perform three-way verification. Narrative-Aware Logistic required a median 3,741.3 ms wall time and 2.798 CPU seconds per 1,000 pairs, giving a 15 W CPU-time proxy of 1.166×10^{-5} kWh (IQR 9.466×10^{-7}). Word+Char required 3,494.4 ms, 2.329 CPU seconds, and 9.703×10^{-6} kWh; the character n-gram dual-output pipeline required 2,517.2 ms, 1.829 CPU seconds, and 7.621×10^{-6} kWh. Among ranking-only systems, BM25 used 178.2 ms, 0.089 CPU seconds, and 3.695×10^{-7} kWh, while LSA-128 used 621.0 ms, 0.440 CPU seconds, and 1.834×10^{-6} kWh. SmolLM2 used a median 58,083.3 ms wall time and 24.702 CPU seconds per 1,000 pairs, giving 1.029×10^{-4} kWh (IQR 1.944×10^{-5}); the full 176-query pass took 111.481 wall seconds and 46.515 CPU seconds.

Table 12. Component final-fit time, warm inference latency, model size, and standardized CPU-energy proxy.

Model	Component Final Fit Wall Seconds	Warm Wall Ms/1,000 [iqr]	Warm CPU Seconds Per 1000 Median	Energy Kwh/1,000 [iqr]	Artifact Size MB	MAP	Macro F1
Word+Char TF-IDF Logistic	1.787	3494.4 [149.8]	2.329	9.70e-06 [9.04e-07]	1.810	0.670	0.392
Word TF-IDF Cosine	—	482.9 [98.2]	0.388	1.62e-06 [1.21e-07]	0.324	0.646	—
BM25	—	178.2 [87.8]	0.089	3.70e-07 [9.33e-08]	0.101	0.649	—
Narrative-Aware Logistic	0.016	3741.3 [1637.8]	2.798	1.17e-05 [9.47e-07]	28.32	0.657	0.394
Compact N-gram LM	0.034	2517.2 [567.9]	1.829	7.62e-06 [5.08e-07]	0.682	0.655	0.387
LSA-128 Cosine	—	621.0 [47.7]	0.440	1.83e-06 [2.59e-07]	27.96	0.665	—
Lexical-Semantic Logistic	0.012	3695.8 [199.7]	2.561	1.07e-05 [6.48e-07]	28.24	0.653	0.371
SmolLM2-135M top-choice + BM25	—	58083.3 [14242.1]	24.70	1.03e-04 [1.94e-05]	100.6	0.647	—

Figure 6 places ranking quality and verification quality in separate Pareto panels because their workloads and outcomes differ. Word+Char occupied the strongest measured ranking position, while the character n-gram classifier was the least costly dual-output model and Narrative-Aware Logistic had the highest macro-F1. Under the deliberately compressed prompt, SmolLM2 consumed approximately 279 times BM25's per-pair CPU-time proxy without improving MAP. The energy numbers are standardized CPU-time estimates rather than meter readings, so they quantify the recorded local workload but do not establish a universal ordering across hardware or model families (Henderson et al., 2020; Upravitelev et al., 2025).

Taken together, the experiments reveal complementary strengths rather than a single dominant system. Sparse Word+Char features gave the best candidate ordering; narrative-aware features shifted recall toward the smaller Refutes class and improved the descriptive calibration and AURC values relative to global scaling. None of the four confirmatory differences was statistically significant after multiplicity adjustment, and neither target-risk screen passed. The evaluation is also bounded to annotated candidate pools and one abstract per decision. Open-corpus retrieval, evidence chains across documents, and claims requiring synthesis remain outside this design (Jiang et al., 2020). Within that boundary, the results provide a coherent basis for routing low-confidence relations to additional search or expert review while reporting the computational cost of each local decision.

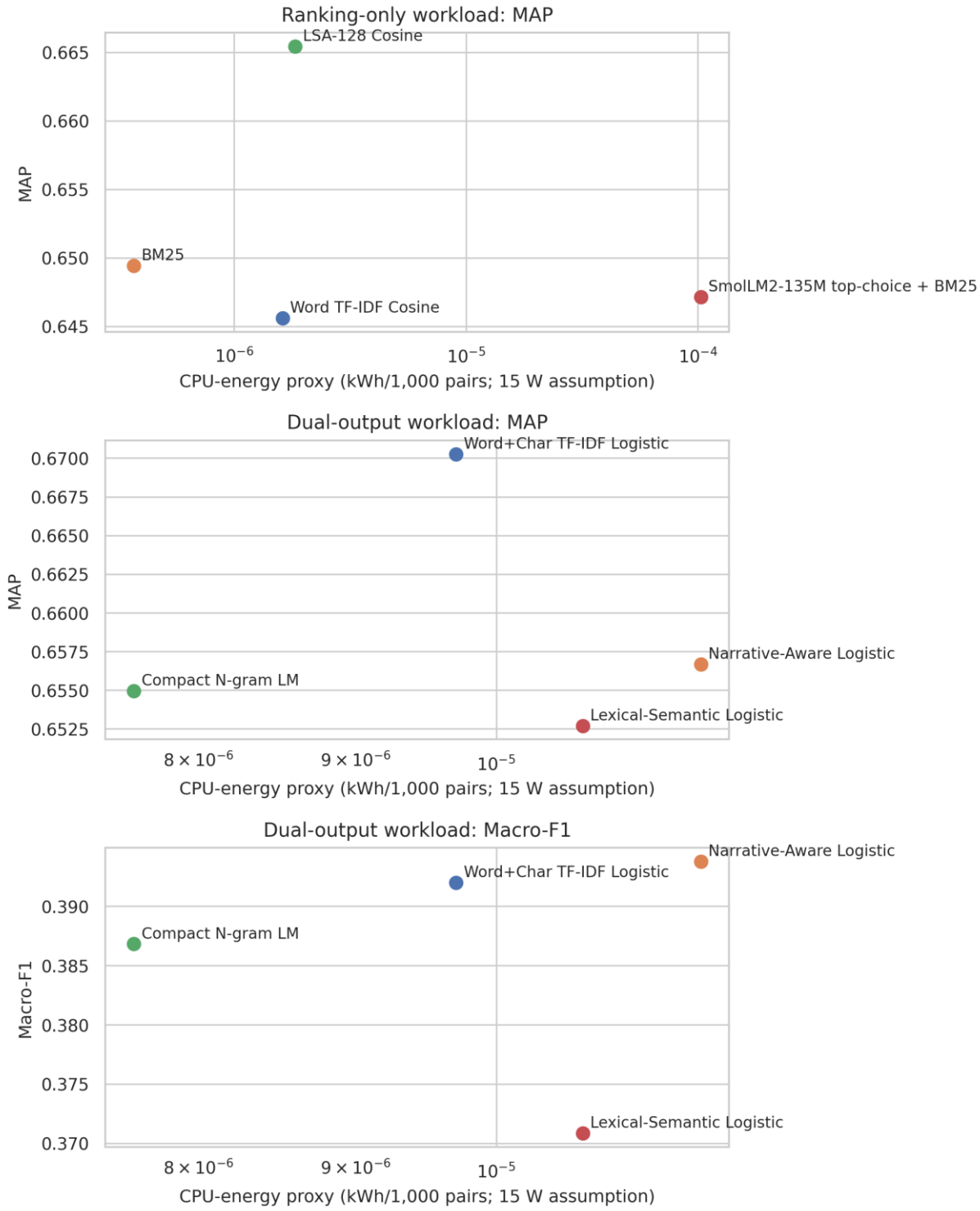

Figure 6. Effectiveness–efficiency frontier under a controlled CPU boundary.

5. Conclusions

This study established a leakage-controlled, resource-conscious evaluation of climate-claim verification on ClimateCheck 2025. It separated evidence ordering within annotated candidate pools from three-way relation classification. Word+Char TF–IDF Logistic led reranking with MAP 0.670, whereas Narrative-Aware Logistic achieved MAP 0.657 and the highest macro-F1, 0.394. The quantized SmolLM2-135M top-choice hybrid obtained MAP 0.647 under its deliberately compressed 128-token prompt and used 1.029×10^{-4} kWh per 1,000 pairs under the standardized CPU-time proxy. Narrative-conditioned temperature scaling produced ECE 0.025 and AURC 0.551, but neither prespecified target-risk screen yielded a feasible operating point and none of the four confirmatory contrasts was significant after Holm adjustment.

The evidence supports a measured trade-off rather than universal narrative-aware superiority. Sparse features ranked candidates most effectively; predicted narrative context altered class balance and reduced descriptive calibration errors relative to global scaling. The narrative-aware dual-output pipeline used 2.798 CPU seconds and 3,741.3 ms wall time per 1,000 pairs, equivalent to a standardized 15 W CPU-time proxy of 1.166×10^{-5} kWh. Future open-corpus and multi-abstract systems should preserve the same separation among retrieval scope, relation accuracy, calibrated uncertainty, abstention feasibility, and computational cost.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Abu Ahmad, R., Usmanova, A., & Rehm, G. (2025a). The ClimateCheck dataset: Mapping social media claims about climate change to corresponding scholarly articles. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)* (pp. 42–56). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sdp-1.5

[2] Abu Ahmad, R., Usmanova, A., & Rehm, G. (2025b). The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)* (pp. 263–275). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sdp-1.24

[3] Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C., & Simonsen, J. G. (2019). MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4685–4697). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1475

[4] 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. *Information Electrical and Electronic Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[5] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronics and Communications Systems, 6*(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533

[6] Ben Allal, L., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Piqueres Lajarín, A., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.-S., Fourrier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., . . . Wolf, T. (2025). SmolLM2: When Smol goes big—Data-centric training of a small language model [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2502.02737

[7] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

[8] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[9] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509

[10] 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

[11] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI Credit Card Clients Dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

[12] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[13] 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

[14] Coan, T. G., Boussalis, C., Cook, J., & Nanko, M. O. (2021). Computer-assisted classification of contrarian claims about climate change. *Scientific Reports, 11*, Article 22320. https://doi.org/10.1038/s41598-021-01714-4

[15] Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE, 12*(5), e0175799. https://doi.org/10.1371/journal.pone.0175799

[16] Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., & Leippold, M. (2020). CLIMATE-FEVER: A dataset for verification of real-world climate claims [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2012.00614

[17] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* (Vol. 30). https://papers.nips.cc/paper/7073-selective-classification-for-deep-neural-networks

[18] 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* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[19] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[20] 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

[21] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[22] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research, 21*(248), 1–43. https://jmlr.org/papers/v21/20-312.html

[23] Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446. https://doi.org/10.1145/582415.582418

[24] Jiang, Y., Bordia, S., Zhong, Z., Dognin, C., Singh, M., & Bansal, M. (2020). HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3441–3460). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.309

[25] Jin, J. (2025a). 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

[26] Jin, J. (2025b). 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

[27] Jin, J., Huang, T., & Lu, S. (2024a). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI Credit Card Default Dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[28] Jin, J., Huang, T., & Lu, S. (2024b). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[29] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[30] Kiepura, A., & Lam, J. (2025). ClimateCheck2025: Multi-stage retrieval meets LLMs for automated scientfic fact-checking. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)* (pp. 293–306). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sdp-1.28

[31] 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

[32] Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science, 8*(12), 2100707. https://doi.org/10.1002/advs.202100707

[33] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[34] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[35] 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

[36] 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

[37] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[38] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[39] 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

[40] 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

[41] 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

[42] 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

[43] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[44] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[45] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[46] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/JACS.2024.40809

[47] 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

[48] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/information-retrieval-book.html

[49] 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

[50] 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

[51] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/JACS.2024.40409

[52] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[53] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

[54] Roelofs, R., Cain, N., Shlens, J., & Mozer, M. C. (2022). Mitigating bias in calibration error estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics* (Vol. 151, pp. 4036–4054). PMLR. https://proceedings.mlr.press/v151/roelofs22a.html

[55] Rojas, C., Algra-Maschio, F., Andrejevic, M., Coan, T., Cook, J., & Li, Y.-F. (2024). Hierarchical machine learning models can identify stimuli of climate change misinformation on social media. *Communications Earth & Environment, 5*, Article 436. https://doi.org/10.1038/s43247-024-01573-7

[56] Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* (pp. 623–632). Association for Computing Machinery. https://doi.org/10.1145/1321440.1321528

[57] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[58] Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

[59] 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

[60] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG Benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[61] 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

[62] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[63] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[64] Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

[65] Upravitelev, M., Duran-Silva, N., Woerle, C., Guarino, G., Mohtaj, S., Yang, J., Solopova, V., & Schmitt, V. (2025). Comparing LLMs and BERT-based classifiers for resource-sensitive claim verification in social media. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)* (pp. 281–287). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sdp-1.26

[66] Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7534–7550). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.609

[67] 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

[68] 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

[69] Wang, J., Chen, K., Chen, Z., He, P., & Zheng, W. (2025). Winning ClimateCheck: A multi-stage system with BM25, BGE-reranker ensembles, and LLM-based analysis for scientific abstract retrieval. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)* (pp. 276–280). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.sdp-1.25

[70] 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

[71] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2). https://doi.org/10.18196/eist.v6i2.31232

[72] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[73] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[74] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[75] Xin, Q. (2026b). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit Dataset (HELOC-motivated). *Journal of Information Technology, 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

[76] Xin, Q. (2026c). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 325–334. https://doi.org/10.28989/avitec.v8i2.3973

[77] Xin, Q. (2026d). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

[78] Xin, Q. (2026e). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. *Journal of Information and Technology, 14*(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

[79] Xin, Q. (2026f). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[80] 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

[81] 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

[82] 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

[83] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[84] 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

[85] 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

[86] Zhang, J. (2025). From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment. *Journal of Technology Informatics and Engineering, 4*(1), 263–283. https://doi.org/10.51903/jtie.v4i1.524

[87] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[88] 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

[89] 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

[90] 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

[91] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92–97. https://doi.org/10.23977/aetp.2026.100213

[92] 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

[93] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545

[94] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[95] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[96] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[97] Zheng, L., Li, C., Zhang, X., Shang, Y.-M., Huang, F., & Jia, H. (2024). Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 9274–9281). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.551

[98] 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

[99] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[100] 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), iaae020. https://doi.org/10.1093/imaiai/iaae020

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

[102] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[103] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[104] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[105] 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

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

[107] 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