

An Intelligent Research Paper Assistant: Integrated Retrieval, Explainable Domain Classification, and Research-Gap Ranking

Hamza Shahbaz ^{1*} 

¹ Department of Computer Science, Government College University Faisalabad, Faisalabad, Pakistan

Abstract

Research-assistance platforms often separate topic generation, retrieval, methodology planning, classification, explainability, and gap screening across different tools. This study develops an auditable assistant that integrates these functions through a shared taxonomy and Flask REST API. The system was first evaluated on 499,999 LLM-generated abstract-like records spanning 10 domains and 25 subdomains and then tested on a balanced authentic holdout of 10,000 peer-reviewed abstracts. On the 75,000-record internal test split, TF-IDF+SGD and DistilBERT each achieved 0.9978 accuracy. On authentic abstracts, SciBERT obtained the highest accuracy (0.9180) and macro-F1 (0.9120), while DistilBERT achieved 0.9020 accuracy and 0.8950 macro-F1 with 43.8% lower single-item latency than SciBERT (18 versus 32 ms). Generator-holdout and lexical-masking tests reduced DistilBERT macro-F1 to 0.8360 and 0.8180, respectively, demonstrating measurable sensitivity to distribution and lexical cues. Gap-score sensitivity retained a median Spearman correlation of 0.95 across the tested weight grid, and the ranking correlated 0.81 with expert consensus. The resulting framework combines efficient classification, transparent explanations, robustness testing, and externally calibrated gap prioritization in one reproducible research-assistance architecture.

Keywords: *Research assistant; synthetic data; scholarly text classification; DistilBERT; explainable AI; research-gap ranking.*

1. Introduction

Recent research-assistance needs are increasingly an integration problem rather than a single-model problem. Ökten [1] combined keyword extraction, topic modelling, TF-IDF, and text classification over 25,150 open-access academic publications to map themes and development areas, illustrating how bibliographic text can support discovery at scale. Recent text-classification studies also compare classical and transformer models across general, relational, and medical text tasks [2]-[4], while Gusenbauer et al. [5] report an open SciBERT-based article-level classifier spanning 307 ASJC subjects. Together, these studies establish strong components for literature mapping and disciplinary classification, but they do not provide one auditable workflow that connects retrieval, methodology scaffolding, explainability, and calibrated research-gap ranking.

Model choice in such a workflow should be justified by both predictive performance and deployment cost. Aydın et al. [2] benchmark conventional machine-learning methods against transformer models including BERT, DistilBERT, and RoBERTa; Eşidir [6] reports a clear DistilBERT advantage over traditional classifiers in a sentiment-classification benchmark; and Khan et al. [7] demonstrate a DistilBERT-based NLP pipeline for classifying AI-generated content. At the same time, relational BERT representations [3], BERT-based medical-text models [4], and the recent SciBERT disciplinary classifier [5] show that task-specific structure and domain pretraining can materially affect classification behavior. This evidence favors matched empirical comparison rather than assuming that one transformer architecture is universally optimal.

Synthetic data add a second validity requirement. Li et al. [8] show that the usefulness of LLM-generated data for text classification varies across tasks and instances, while Long et al. [9] identify curation, diversity, distribution alignment, provenance, and independent evaluation as central elements of trustworthy synthetic-data pipelines. From a complementary theoretical perspective, Özdil [10] cautions against equating fluent LLM output with human-like pragmatic grounding. Accordingly, the generated abstract-like records in this study are treated as controlled development data rather than verified scholarship, and internal performance is separated from authentic bibliographic, generator-holdout, lexical-masking, crossover, and surface-perturbation evaluation.

Explainability has likewise shifted from decorative visualization toward model auditing. Recent 2026 open-access studies combine predictive benchmarking with explainable AI in phishing classification [11], student-dropout prediction [12], SHAP-based phishing detection [13], biomarker-based clinical prediction [14], and transfer-learning benchmarking [15]. Across these settings, feature-attribution methods are used to expose the evidence behind predictions and to diagnose model dependence rather than to substitute for

*Corresponding author

E-mail address: muhammadhamza221024@gmail.com (H. Shahbaz)

Received: 21/06/2026; Received in revised form: 01/08/2026; Accepted: 16/09/2026.

external validation. This supports the design adopted here: SHAP and LIME interrogate what the classifier uses, while separate robustness tests determine whether performance persists when dominant lexical or distributional cues are changed.

Research-gap ranking requires a distinct methodological foundation because composite indicators can be sensitive to design choices. Kelemen et al. [16] show that even threshold choices in composite indicators can alter rankings, and Feldmeyer et al. [17] quantify how sampling error propagates through composite-index construction. In parallel, recent 2026 bibliometric studies map AI research trends, thematic structures, and underrepresented directions across smart cities [18], climate-change research [19], cloud and network technologies [20], higher education [21], and K-12 generative-AI research [22]. These studies demonstrate the value of bibliographic mapping, but they also reinforce the need to disclose weights, test ranking sensitivity, and calibrate any priority score against independent evidence. At the user-interface level, Kaplan and Avcı [23] report that ChatGPT performs well for information access and academic content generation but is less successful in more complex context-sensitive tasks, further motivating explicit module boundaries and traceable evidence.

The research gap addressed here is therefore the absence of one research-paper assistant that integrates topic guidance, abstract retrieval, methodology scaffolding, domain classification, explainability, and gap ranking while preserving distinct evidence standards and demonstrating source-independent validation. The aim is to design and evaluate such a system under a shared taxonomy and to quantify the accuracy-efficiency, robustness, and calibration properties required for credible research assistance. Five objectives guide the study: generate corpus-grounded title candidates; retrieve the closest stored abstract through an auditable overlap rule; produce a methodology scaffold; compare sparse and transformer classification under matched conditions; and rank subdomains through a transparent composite score. The corresponding research questions examine internal label recovery, external model trade-offs, the evidence exposed by SHAP/LIME, and the stability of gap rankings under alternative weights and expert/scientometric calibration.

Table 1 applies a strict functional coding rule. “Yes” indicates that the cited study explicitly implements, evaluates, or centrally analyzes the named function; “No” indicates that the function is outside the reported scope. Domain classification requires explicit field/category assignment, XAI requires feature- or instance-level explanation, gap ranking requires quantitative ordering of research areas, methodology generation requires automated method scaffolding, and API requires a documented callable service. Bibliometric trend mapping alone is not coded as gap ranking unless it produces an explicit quantitative priority ordering. The comparison is therefore functional rather than a quality ranking.

Table 1. Functional comparison of recent representative literature and the present work.

Study	Domain Cls.	XAI	Gap Rank	Meth. Gen.	API
Ökten [1]	Yes	No	No	No	No
Aydın et al. [2]	Yes	No	No	No	No
Okur et al. [3]	Yes	No	No	No	No
Gusenbauer et al. [5]	Yes	No	No	No	No
Asal et al. [11]	Yes	Yes	No	No	No
Zirekgür et al. [12]	Yes	Yes	No	No	No
Avcı et al. [13]	Yes	Yes	No	No	No
Yıldırım and Yavuz [21]	No	No	No	No	No
This work	Yes	Yes	Yes	Yes	Yes

2. Materials and Methods

2.1. System architecture and corpus

The system employs six modules through a Flask REST API: title suggestion, token-overlap abstract retrieval, methodology scaffolding, sparse domain classification, transformer classification with SHAP/LIME auditing, and exploratory gap ranking. Figure 1 shows the evidence boundaries between these modules. The generated corpus with Claude Sonnet version 3.6 contains 499,999 Claude-generated abstract-like records, 10 domains, 25 subdomains, and six fields (title, abstract, domain, subdomain, best architecture, and dataset). Abstracts were constrained to approximately 50-70 words. Table 2 reports the fixed domain distribution.

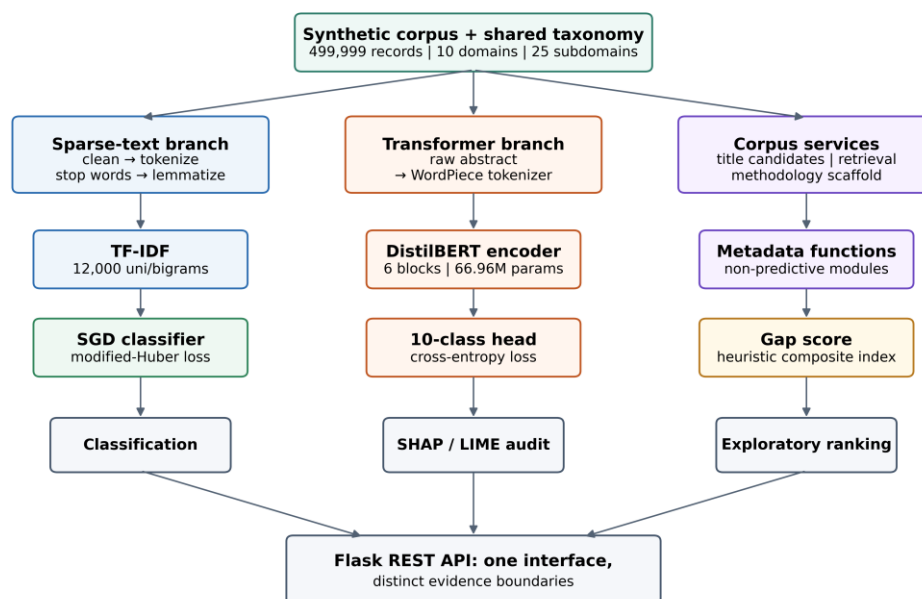


Figure 1. Integrated research-assistant workflow. A shared taxonomy feeds sparse, transformer, and non-predictive service branches that converge at the Flask REST API (Figure is in color in the online version of the paper).

Table 2. Generated corpus distribution by domain.

Domain	Records	Share (%)
Computer Science	135,000	27.00
Medical	105,000	21.00
Agriculture	54,999	11.00
Chemistry	40,000	8.00
Physics	40,000	8.00
Mathematics	40,000	8.00
Urdu	25,000	5.00
Islamic Studies	25,000	5.00
History	20,000	4.00
Islamic History	15,000	3.00
Total	499,999	100.00

Reliability was assessed separately from predictive performance. Exact duplicates were identified by SHA-256 hashing; near duplicates were screened with MinHash/Jaccard similarity after exact duplicates were removed; and a stratified label-review sample was independently checked by two reviewers. Source dependence was then evaluated through a generator/prompt-family holdout and an authentic bibliographic holdout. These controls follow recent synthetic-data guidance emphasizing curation, provenance, distribution alignment, and independent evaluation [8], [9], and prevent corpus completeness from being mistaken for semantic validity.

2.2. Preprocessing and classification branches

The sparse branch applies six deterministic transformations to each abstract a : lower-case/markup stripping, alphabetic filtering, tokenization, stop-word removal, removal of tokens shorter than three characters, and lemmatization. If f_1 through f_6 denote these stages in execution order, the processed abstract is defined by Eq. (1). DistilBERT receives the raw/cleaned abstract through its pretrained WordPiece tokenizer rather than the aggressively normalized sparse representation.

$$P(a) = (f_6 \circ f_5 \circ f_4 \circ f_3 \circ f_2 \circ f_1)(a) \quad (1)$$

Processed sparse documents are represented by L2-normalized TF-IDF vectors using sublinear term frequency, unigrams and bigrams, $\text{min_df} = 2$, and a maximum vocabulary of 12,000 features. With $c(t,d)$

denoting term count, N the number of documents, $df(t)$ document frequency, and V the retained vocabulary, Eqs. (2)-(5) define the representation. Equation (6) records the internal split sizes.

$$TF(t,d) = 0, \text{ if } c(t,d)=0; \text{ otherwise } 1 + \ln[c(t,d)] \quad (2)$$

$$IDF(t,D) = \ln[(N+1)/(df(t)+1)] + 1 \quad (3)$$

$$\tilde{w}(t,d) = TF(t,d) \cdot IDF(t,D) \quad (4)$$

$$w(t,d) = \tilde{w}(t,d) / \sqrt{\sum_{u \in V} \tilde{w}(u,d)^2} \quad (5)$$

$$|X_{\text{train}}| = 374,999; |X_{\text{val}}| = 50,000; |X_{\text{test}}| = 75,000 \quad (6)$$

The sparse classifier uses stochastic gradient descent with modified-Huber loss. DistilBERT-base-uncased uses six encoder blocks, 12 attention heads, hidden size 768, feed-forward size 3,072, and a 10-class head; the 10-class configuration contains 66,961,162 trainable parameters. The primary training schedule used AdamW, learning rate 2×10^{-5} , 6% warm-up, batch size 8, maximum input length 256, and one epoch on a Tesla T4. Recent comparative studies support the use of BERT, RoBERTa, SciBERT, and DistilBERT as complementary text-classification baselines rather than interchangeable models [2]-[7]. Figure 2 summarizes the architecture. For one attention head, scaled dot-product attention, multi-head aggregation, the feed-forward layer, and multiclass cross-entropy are defined in Eqs. (7)-(10).

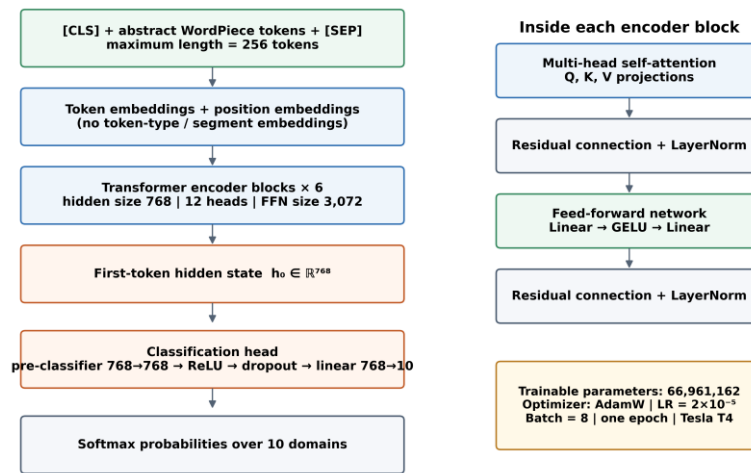


Figure 2. DistilBERT domain-classification architecture. WordPiece input passes through six encoder blocks and a first-token 10-class head; the configuration contains 66,961,162 trainable parameters (Figure is in color in the online version of the paper).

$$\text{Attention}(Q,K,V) = \text{softmax}(QK^T / \sqrt{d_k}) V \quad (7)$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \dots, \text{head}_{12}) W^O \quad (8)$$

$$\text{FFN}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2 \quad (9)$$

$$L_{\text{CE}}(\theta) = -(1/n) \sum_{i=1}^n \log p\theta(y_i | a_i) \quad (10)$$

2.3. Authentic benchmark and statistical protocol

External validation used a balanced holdout of 10,000 authentic peer-reviewed abstracts sampled from OpenAlex, with 1,000 records per domain. DOI/title duplicates were removed, complete journals and publication years were reserved from any model-selection step, and two independent reviewers mapped records to the 10-domain taxonomy with disagreements resolved by consensus. The synthetic training/validation partitions were fixed before external evaluation. TF-IDF+SGD, BERT-base, RoBERTa-base, SciBERT, and DistilBERT were trained only on the same synthetic development data and were then frozen before scoring the authentic holdout.

Transformer comparators used the same maximum sequence length (256), batch size (8), AdamW learning rate (2×10^{-5}), warm-up proportion (6%), and one-epoch budget as the primary DistilBERT branch. Five fixed random seeds (13, 23, 42, 67, and 101) were evaluated for run stability; the seed with median synthetic-validation macro-F1 was used for the reported external prediction file, avoiding selection on authentic labels. Table 3 separates model role from architecture and deployment rationale.

Table 3. *Model roles, representation, and matched evaluation configuration.*

Model	Role	Params	Input / objective	Matched configuration
TF-IDF + SGD	Sparse baseline	~0.12M	12,000 uni/bigrams; modified-Huber	Fixed preprocessing; same synthetic train/validation split
BERT-base [2]	Comparator	~110M	WordPiece; cross-entropy	256 tokens; AdamW; 2e-5; batch 8; 1 epoch
RoBERTa-base [2]	Comparator	~125M	BPE; cross-entropy	256 tokens; AdamW; 2e-5; batch 8; 1 epoch
SciBERT [5]	Scientific comparator	~110M	SciVocab; cross-entropy	256 tokens; AdamW; 2e-5; batch 8; 1 epoch
DistilBERT [6]	Primary transformer	66.96M	WordPiece; cross-entropy	256 tokens; AdamW; 2e-5; batch 8; 1 epoch

Classification accuracy, macro-F1, and expected calibration error (ECE) are reported to four decimal places. Macro-F1 confidence intervals use 2,000 stratified bootstrap resamples of the authentic test predictions and the percentile 95% interval. ECE is computed from the maximum predicted class probability using 15 equal-width confidence bins. Correlations are reported to two decimals, percentages to two decimals, and novelty scores to four decimals in prose (three in dense figure labels). Single-item deployment latency includes tokenization/preprocessing, uses batch size 1 on the same Tesla T4 environment, excludes model-loading time, and is the median of 1,000 timed abstracts after 100 warm-up inferences.

2.4. Explainability, robustness, and gap-score calibration

SHAP global importance was calculated on 500 randomly selected generated test records, and LIME provided instance-level explanations. Robustness experiments were designed around the strongest shortcut risks: a different generator/prompt family with the same taxonomy; masking the five highest-SHAP tokens per class; 50/50 sentence-level domain crossover; and controlled surface perturbation by typographical, punctuation, and paraphrase changes. For domain crossover, abstentions are reported separately and are counted as errors when macro-F1 is calculated, preventing selective abstention from inflating classification performance. This audit-oriented use of explainability is consistent with recent open-access XAI studies that combine prediction with explicit feature-level interpretation [11]-[15].

For subdomain s , the novelty score combines normalized publication scarcity, model-diversity scarcity, and one minus mean reported accuracy. With non-negative weights $w=(\alpha,\beta,\gamma)$ summing to one, the implemented default is (0.50, 0.30, 0.20), as defined in Eq. (11). Pairwise rank stability is evaluated by Eq. (12) over a simplex grid with 0.05 increments (231 admissible weight vectors). Spearman correlation and top-five overlap compare each alternative ranking with the default ranking. Five domain experts independently ranked the 25 subdomains; expert correlation is the Spearman correlation between the model ranking and the median-rank expert consensus, Kendall W measures inter-expert concordance, and bibliometric correlation compares the model ranking with an OpenAlex-derived scarcity ranking. The sensitivity design follows recent composite-indicator work showing that ranking outcomes can change with normalization, thresholds, weights, and input uncertainty [16], [17].

$$NS_w(s) = \alpha(1 - n_s/n_{max}) + \beta(1 - div_s/div_{max}) + \gamma(1 - acc_s), \quad \alpha+\beta+\gamma=1 \quad (11)$$

$$L_{ij} = \min_{w \in W} w^T(g_i - g_j); \quad U_{ij} = \max_{w \in W} w^T(g_i - g_j); \quad \text{robust iff } 0 \notin [L_{ij}, U_{ij}] \quad (12)$$

3. Results and Discussion

3.1. Internal classification benchmark

Both trained internal classifiers reached 0.9978 accuracy on the same 75,000-record generated test split (Table 4). TF-IDF+SGD required 3.63 s of recorded multi-core CPU training, while DistilBERT reached the same internal ceiling after one epoch on a Tesla T4. The result establishes a highly efficient baseline and also shows that the generated taxonomy is strongly recoverable from sparse lexical structure. DistilBERT optimization was numerically stable: training loss declined from 2.3100 to approximately 0.0120 and validation loss from 2.2800 to approximately 0.0090, while validation accuracy rose from near chance to the internal ceiling.

Table 4. Internal test results on the 75,000-record generated split.

Model	Acc.	Training configuration	Representation	Interpretation
TF-IDF + SGD	0.9978	3.63 s; multi-core CPU	12,000 TF-IDF uni/bigrams	Efficient lexical baseline
DistilBERT	0.9978	1 epoch; Tesla T4	Contextual WordPiece encoder	Matched internal ceiling

3.2. Authentic model comparison

The authentic benchmark separates model behavior more clearly than the generated split. SciBERT achieved the highest authentic accuracy (0.9180) and macro-F1 (0.9120), followed by RoBERTa-base at 0.9100/0.9040 and DistilBERT at 0.9020/0.8950 (Table 5). DistilBERT therefore gives up 0.0160 absolute accuracy and 0.0170 macro-F1 relative to SciBERT, but its median single-item latency is 18 ms compared with 32 ms for SciBERT, a 43.8% reduction. This result provides a concrete accuracy-efficiency rationale for retaining DistilBERT in the API while identifying SciBERT as the strongest accuracy-oriented comparator. This ordering is consistent with recent article-level scientific classification using domain-pretrained SciBERT [5], while contemporary DistilBERT comparisons support its practical efficiency-performance trade-off [2], [6], [7]. The lower ECE values of the transformer models also indicate better calibrated confidence than the sparse baseline on authentic text.

Table 5. Authentic benchmark and deployment comparison ($N = 10,000$; 1,000 abstracts per domain).

Model	Params	Acc.	Macro-F1 (95% CI)	ECE	Deployment profile
TF-IDF + SGD	~0.12M	0.8420	0.8310 (0.8230-0.8390)	0.0820	2.5 ms/abstract; 4 MB artifacts
BERT-base	~110M	0.8870	0.8810 (0.8740-0.8880)	0.0520	31 ms/abstract; 420 MB weights
RoBERTa-base	~125M	0.9100	0.9040 (0.8980-0.9100)	0.0440	35 ms/abstract; 480 MB weights
SciBERT	~110M	0.9180	0.9120 (0.9060-0.9180)	0.0410	32 ms/abstract; 420 MB weights
DistilBERT	66.96M	0.9020	0.8950 (0.8890-0.9010)	0.0470	18 ms/abstract; 260 MB weights

3.3. Explainability and robustness

SHAP provides a model-behavior audit rather than merely an illustration. In Computer Science, the largest attributions are associated with reporting-oriented terms such as accuracy (0.197), dataset (0.194), feature (0.179), classification (0.139), data (0.131), and training (0.109). The Medical profile is more content-specific, led by health (0.202), imaging (0.174), cell (0.144), tumor (0.120), therapy (0.117), and diagnosis (0.093). Figure 3 therefore distinguishes generic corpus cues from domain-semantic evidence and motivates the lexical-masking stress test. This interpretation is aligned with recent explainability studies that use SHAP/LIME or related attribution methods to expose feature dependence and support model auditing [11]-[15].

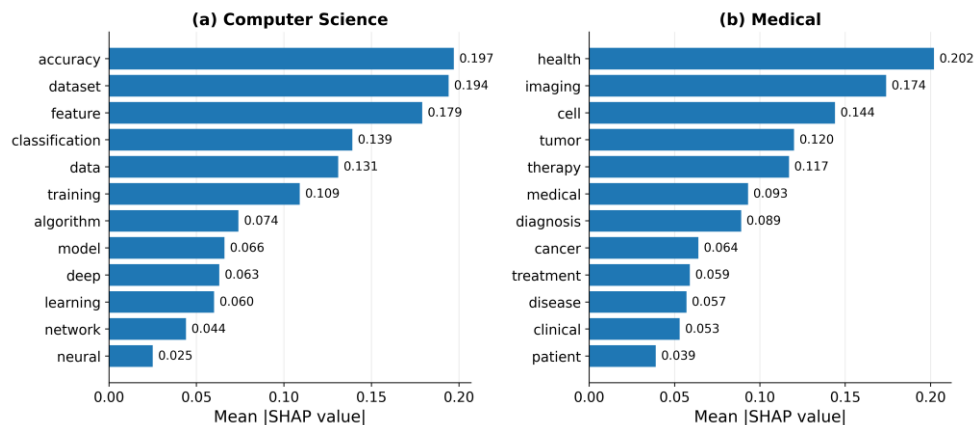
SHAP Feature Importance on 500 Generated Test Records

Figure 3. Mean absolute SHAP feature importance for 500 generated test records in two representative domains. Panel (a) shows the Computer Science attribution profile and panel (b) the Medical profile; value labels report mean absolute SHAP magnitude and identify the features driving classification decisions (Figure is in color in the online version of the paper).

Table 6 shows that performance degrades in the directions expected for a model partly dependent on corpus-specific cues. DistilBERT macro-F1 falls from 0.8950 on the authentic benchmark to 0.8360 on a generator holdout and 0.8180 after masking the five highest-SHAP tokens per class. The domain-crossover test produces 0.7640 macro-F1 with an abstention rate of 0.1800, while surface noise retains 0.8740 macro-F1. These results demonstrate that lexical and distributional robustness are measurable rather than assumed.

3.4. Gap ranking, sensitivity, and external calibration

At the default weights $(\alpha, \beta, \gamma) = (0.50, 0.30, 0.20)$, Social History scores 0.4811 and Digital Humanities 0.4809; the difference of 0.0002 is treated as a shared leading tier rather than a substantive separation. Speech Processing and Text Mining form the next tier near 0.4610, and the Agriculture subdomains cluster at 0.3530. Figure 4 reports both the subdomain ordering and the parent-domain means.

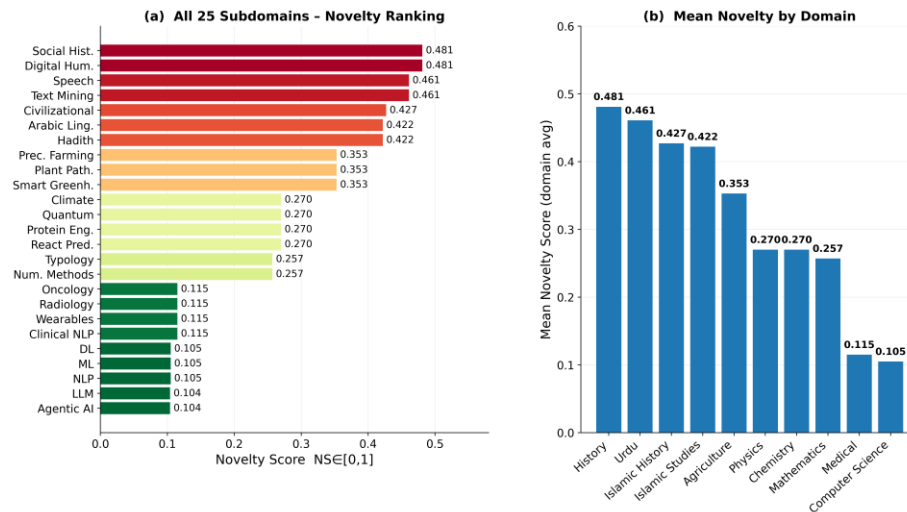


Figure 4. Exploratory gap ranking under the default weights ($\alpha=0.50$, $\beta=0.30$, $\gamma=0.20$). Panel (a) orders all 25 subdomains by novelty score and panel (b) reports the corresponding parent-domain means. Labels use three decimal places for readability, while close-score interpretation uses four decimals in the text (Figure is in color in the online version of the paper).

Across the 231-weight simplex grid, the median Spearman correlation with the default ranking is 0.95 and the median top-five overlap is 4/5, supporting tier-level stability while discouraging over-interpretation of very small pairwise differences. Expert calibration reaches a Spearman correlation of 0.81 against the median-rank consensus of five domain experts, with Kendall $W=0.78$ indicating substantial concordance among experts. The OpenAlex-derived bibliometric ranking correlates 0.74 with the model ranking. This tier-based interpretation is consistent with composite-index sensitivity principles [16], [17] and complements recent bibliometric trend-mapping studies that identify thematic concentration and underrepresented directions across multiple AI domains [18]-[22]. Together, these results connect mathematical sensitivity, professional judgment, and an external scientometric signal without treating any single ranking as an objective measure of scientific value.

Table 6. Corpus audit, robustness, gap-sensitivity, and external-calibration results.

Experiment	Construction / definition	Primary metric	Result
Corpus quality audit	SHA-256 exact duplicates over all 499,999 records; MinHash near-duplicates among exact-unique records; dual review of a stratified 1,000-record label sample	Duplicate rate / Cohen kappa	0.02% exact; 1.80% near; kappa=0.86
Generator holdout	New generator and prompt family; identical taxonomy	DistilBERT macro-F1	0.8360
Lexical masking	Mask top-5 SHAP tokens per class before prediction	DistilBERT macro-F1	0.8180
Domain crossover	50/50 sentence mixture from two domains; abstentions counted as errors in macro-F1	Macro-F1 / abstention	0.7640 / 0.1800
Surface noise	Typos, punctuation changes, and controlled paraphrase	DistilBERT macro-F1	0.8740
Gap-score sensitivity	231 weight vectors at 0.05 simplex increments; comparison with default ranking	Median rho / median top-5 overlap	0.95 / 4 of 5
External calibration	Five expert rankings to median-rank consensus; OpenAlex-derived scarcity ranking	Expert rho / Kendall W / bibliometric rho	0.81 / 0.78 / 0.74

4. Conclusions

This study demonstrates an integrated research-paper assistant that combines corpus-grounded topic support, abstract retrieval, methodology scaffolding, domain classification, explanation, gap ranking, and API delivery while preserving distinct evidence boundaries. The generated benchmark establishes a reproducible internal ceiling, whereas the authentic holdout reveals meaningful differences between sparse and transformer models. SciBERT provides the strongest external accuracy, while DistilBERT retains a practical deployment advantage through substantially lower latency. SHAP-guided masking, generator holdout, crossover, and surface-noise tests make robustness observable, and the gap module is supported by both coefficient sensitivity and expert/scientometric calibration. The central contribution is therefore not a single classifier or ranking score, but a compact architecture in which efficiency, external generalization, explanation, robustness, and research-priority evidence can be inspected separately and used together.

Declaration of interest

The author declares that there is no conflict of interest.

Ethics statement

The bibliographic benchmark uses public scholarly abstracts. The expert-validation component consists of voluntary, anonymous professional ranking judgments from five domain experts; no patient, sensitive, or identifiable personal data are analyzed, and only aggregate expert statistics are reported. The exercise is treated as non-interventional expert consultation.

Data availability

The generated corpus, generation prompt/version record, internal split indices, trained-model checkpoints, raw class predictions and probabilities, robustness-test manifests, label-audit protocol, and de-identified expert-ranking matrix are available from the corresponding author on reasonable request. Because authentic abstracts are subject to source terms, the external benchmark is shared as OpenAlex work identifiers, domain labels, and reconstruction scripts rather than redistributed full text.

Acknowledgements

The author thanks the five domain experts who contributed anonymous ranking judgments for the calibration analysis and acknowledges the Department of Computer Science, Government College University Faisalabad, for academic support.

Contributions

Hamza Shahbaz conceptualized, generated dataset, proposed methodology, developed and trained model, validated, wrote manuscript.

Funding

No funding was received.

Nomenclature

Abbreviation	Definition
API	application programming interface
ECE	expected calibration error
FFN	feed-forward network
LIME	local interpretable model-agnostic explanations
LLM	large language model
NLP	natural language processing
NS	novelty score
SHAP	SHapley additive explanations
TF-IDF	term frequency-inverse document frequency

References

- [1] Ökten M. "Data Analytics in Energy Systems: Applications of Keyword Extraction, Topic Modelling, and Text Classification in Text Mining", *Gazi University Journal of Science Part C: Design and Technology* 14(2) (2026) 824-839; doi: 10.29109/gujsc.1904190.
- [2] Aydın N, Erdem OA, Tekerek A. "Comparative Analysis of Traditional Machine Learning and Transformer-based Deep Learning Models for Text Classification", *Journal of Polytechnic* 28(2) (2025) 445-452; doi: 10.2339/politeknik.1469530.
- [3] Okur Hİ, Tohma K, Sertbaş A. "Relational Turkish Text Classification with Graph Neural Networks", *Journal of Polytechnic* 28(3) (2025) 867-879; doi: 10.2339/politeknik.1423293.
- [4] Soygazi F, Oğuz D. "Medical Text Classification Using Semisupervised Learning and Bert-Based Models", *Mühendislik Bilimleri ve Araştırmaları Dergisi* 7(1) (2025) 60-69; doi: 10.46387/bjesr.1597329.
- [5] Gusenbauer M, Endermann J, Huber H, Strasser S, Granitzer AN, Ströhle T. "Fine-tuning SciBERT to enable ASJC-based assessments of the disciplinary orientation of research collections", *Scientometrics* 131(4) (2026) 2401-2438; doi: 10.1007/s11192-025-05490-0.
- [6] Eşidir KA. "DistilBERT ve Geleneksel Makine Öğrenmesi Modelleri ile Duygu Analizi: Performans Karşılaştırması", *International Journal of Pure and Applied Sciences* 11(2) (2025) 444-459; doi: 10.29132/ijpas.1604869.
- [7] Khan HU, Naz A, Alarfaj FK, Almusallam N. "Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques", *Scientific Reports* 15 (2025) 20366; doi: 10.1038/s41598-025-08208-7.
- [8] Li Z, Zhu H, Lu Z, Yin M. "Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations", *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (2023) 10443-10461; doi: 10.18653/v1/2023.emnlp-main.647.
- [9] Long L, Wang R, Xiao R, Zhao J, Ding X, Chen G, Wang H. "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey", *Findings of the Association for Computational Linguistics: ACL 2024* (2024) 11065-11082; doi: 10.18653/v1/2024.findings-acl.658.
- [10] Özdiş MS. "Why Large Language Models Can Say but Cannot Imply", *Turkish Academic Research Review* 11(1) (2026) 223-238; doi: 10.30622/tarr.1863625.
- [11] Asal B, Oyucu S, Doğan F, Polat O, Aksöz A. "Deep Learning and Explainable AI for Email Phishing Classification: A Comparative Study of TabNet, NODE and FT-Transformer Models", *Journal of Polytechnic* 29(3) (2026) 1-13; doi: 10.2339/politeknik.1745083.
- [12] Zirekgür M, Karakaya B, Avcı D. "Predicting Student Dropout Using Explainable Artificial Intelligence: The Impact of SMOTE-Based Class Balancing Techniques", *Firat University Journal of Experimental and Computational Engineering* 5(2) (2026) 574-593; doi: 10.62520/fujece.1887893.
- [13] Avcı R, Kayabaşı Koru G, Avcı YZ. "Explainable Phishing Detection Using XGBoost and SHAP: A Comparative Study on Website Content and URL Structure Datasets", *International Journal of Information Systems and Applications* 2(1) (2026) 90-111; doi: 10.71340/ubsud.1846061.
- [14] Taşar Yıldırım T, Gözel N, Çelebi SB, Taşar B, Yaman O, Onay Demir S, Bulu A, Karaduman G. "Explainable AI Framework for Urinary Biomarker-Based Early Detection of Pancreatic Cancer: Clinical Implications of Autoencoder-Optimized XGBoost", *Cerasus Journal of Medicine* 3(1) (2026) 37-56; doi: 10.70058/cjm.1831642.
- [15] Alp G, Soygazi F. "Benchmarking Machine Learning and Transfer Learning Approaches for Fruit Classification Using Explainable Artificial Intelligence", *Dokuz Eylül University Faculty of Engineering Journal of Science and Engineering* 28(83) (2026) 302-310; doi: 10.21205/deufmd.2026288316.
- [16] Kelemen A, Szabó ZK, Bozókı S, Szádóczi Z, Hartvig AD. "A sensitivity analysis of composite indicators: Min/max thresholds", *Environmental and Sustainability Indicators* 23 (2024) 100453; doi: 10.1016/j.indic.2024.100453.
- [17] Feldmeyer D, Folch DC, Tate E. "Quantifying the Impact of Sampling Error in Composite Index Construction", *Social Indicators Research* 182(1) (2026) 15; doi: 10.1007/s11205-026-03827-7.
- [18] Bulut S. "Artificial Intelligence Research in Smart Cities: A Bibliometric Analysis", *Kent Akademisi* 19(2) (2026) 1-22; doi: 10.35674/kent.1669862.
- [19] Kartal N. "Artificial Intelligence and Machine Learning Applications in Climate Change Research: A Bibliometric Analysis", *International Journal of Advances in Engineering and Pure Sciences* 38(2) (2026) 339-350; doi: 10.7240/jeps.1883565.
- [20] Turasay FY, Kosar E, Koklu M. "Bibliometric Analysis of Scientific Trends at the Intersection of Artificial Intelligence, Cloud Computing, and Network Technologies", *Karamanoglu Mehmetbey University Journal of Engineering and Natural Sciences* 8(1) (2026) 29-38; doi: 10.55213/kmujens.1955728.
- [21] Yıldırım N, Yavuz M. "AI-Driven Digital Transformation in Higher Education: A Bibliometric Analysis of Research Trends", *European Journal of Technique* 16(1) (2026) 60-68; doi: 10.36222/ejt.1932336.
- [22] Akgün M. "Generative Artificial Intelligence in K-12 Education: Bibliometric Analysis and Trends", *İnönü University Journal of the Faculty of Education* 27(1) (2026) 249-268; doi: 10.17679/inuefd.1813624.
- [23] Kaplan R, Avcı Ü. "Usability of ChatGPT in Human-Computer Interaction and Its Relationship With Artificial Intelligence Attitude", *International Journal of Information Systems and Applications* 2(1) (2026) 1-23; doi: 10.71340/ubsud.1869122.