ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/rajat-kumar-sahoo/661

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 9 & Iss: 1

Research Article

An LLM-Augmented ML Framework for Cross-Domain Sentiment Analysis

Rajat Kumar Sahoo[1]*, Akhilesh Das Gupta[2] and Guru Gobind Singh[3]

[1]Department of Artificial Intelligence and Data Science, India

[2]Institute of Professional Studies, India

[3]Indraprastha University, India

Citation: Rajat KS, Gupta AD, Singh GG. An LLM-Augmented ML Framework for Cross-Domain Sentiment Analysis. *J Artif Intell Mach Learn & Data Sci 2026* 9(1), 3286-3291. DOI: doi.org/10.51219/JAIMLD/rajat-kumar-sahoo/661

Received: 20 November, 2025; **Accepted:** 18 February, 2026; **Published:** 20 February, 2026

***Corresponding author:** Rajat Kumar Sahoo, Department of Artificial Intelligence and Data Science, Email: rajatsahoo2118@gmail.com

ABSTRACT

This paper presents a novel LLM-augmented machine learning framework for cross-domain sentiment analysis that combines traditional ML approaches with large language model assistance. The proposed framework integrates TF-IDF feature extraction, ensemble classification methods (SVM, Random Forest, Gradient Boosting) and dimensionality reduction techniques (LSA, LDA) to achieve competitive performance while maintaining superior computational efficiency and interpretability. Evaluated across three heterogeneous domains-electronics, food & beverage and apparel reviews-the framework achieves 83.7% accuracy with only 2.7% cross-domain degradation. Key innovations include transparent LLM integration for research augmentation, weighted ensemble voting mechanisms and systematic hyperparameter optimization via GridSearchCV. The framework demonstrates practical viability for resource constrained environments, achieving 20-50× faster inference (2.1ms vs 45-120ms) and 8-10× smaller model size compared to deep learning alternatives, while maintaining explainability crucial for regulated domains.

Index Terms: Machine Learning, Sentiment Analysis, TF-IDF, Cross-Domain Transfer Learning, Ensemble Methods, LLM Integration, Natural Language Processing

1. Introduction

The exponential growth of unstructured textual data across digital platforms has created unprecedented demand for automated semantic analysis systems capable of extracting meaningful insights from diverse linguistic contexts. Organizations generate terabytes of text daily through customer reviews, social media interactions and transactional communications, necessitating scalable, efficient and interpretable solutions.

A. Background and motivation

Traditional sentiment analysis approaches relied on manual annotation by domain experts-an approach that is prohibitively expensive and non-scalable at contemporary data volumes. While deep learning architectures, particularly transformers and

large language models, have revolutionized NLP, they demand substantial computational resources, extensive labelled datasets and sophisticated infrastructure for deployment.

This research investigates an alternative paradigm: properly engineered classical machine learning techniques that achieve competitive accuracy while offering profound advantages in computational efficiency, interpretability and cross-domain generalization. The framework uniquely integrates LLM assistance (ChatGPT and Perplexity AI) transparently for literature synthesis and technical debugging, establishing a reproducible methodology for contemporary research practices.

B. Research contributions

The primary contributions of this work include:

- A comprehensive LLM-augmented ML framework combining TF-IDF, LSA, LDA and ensemble methods for semantic analysis.
- Systematic evaluation demonstrating 83.7% accuracy with 2.7% cross-domain degradation across three domains.
- Quantitative comparison revealing 20-50× inference speedup and 8-10× memory reduction versus deep learning.
- Transparent integration methodology for LLM research assistance tools.
- Actionable recommendations for practitioners balancing accuracy, efficiency and interpretability.

C. Paper organization

The remainder of this paper is organized as follows: Section II reviews related work and establishes theoretical foundations. Section III presents the proposed methodology and framework architecture. Section IV details experimental setup and datasets. Section V presents comprehensive results and analysis. Section VI discusses implications and comparisons with alternative approaches. Section VII concludes with limitations and future directions.

2. Related Work and Theoretical Foundations

A. Semantic analysis paradigms

Semantic analysis encompasses automated systems designed to extract, represent and reason about meaning in natural language text[4]. Contemporary approaches employ two primary paradigms: the statistical paradigm models meaning through distributional hypothesis, while the neural paradigm grounds meaning in learned continuous representations.

B. Feature extraction techniques

TF-IDF: Term Frequency-Inverse Document Frequency remains widely deployed for text classification with extensive empirical validation[1,2]. TF-IDF quantifies term importance by combining term frequency within documents and inverse document frequency across corpus:

$$\text{TF-IDF(t,d)} = \log(1 + \text{count}(t,d)) \times \log\left(\frac{N}{\text{df}(t)}\right) \quad (1)$$

where N represents total documents and df(t) is document frequency of term t.

- **Latent semantic analysis:** LSA addresses TF-IDF limitations by applying Singular Value Decomposition to discover latent semantic structure[6]:

$$A \approx U\Sigma V^T \quad (2)$$

where A is the m × n term-document matrix, truncated to k dimensions (k = 50 − 100) to capture essential semantics.

Latent dirichlet allocation: LDA provides probabilistic topic modelling, treating documents as mixtures of latent topics[5]:

$$P(d) = \int P(\theta_d) \prod_w P(w|z,\beta) P(z|\theta_d) d\theta_d \quad (3)$$

C. Classification algorithms

- **Support vector machines:** SVMs find optimal decision boundaries by maximizing margin between classes. For multiclass problems, one-versus-rest decomposition trains k binary classifiers.
- **Random forest:** Random Forest aggregates predictions across hundreds of decision trees trained on random data subsamples[3]. Empirical results demonstrate 84.99% accuracy on anxiety detection and 98.6% on large-scale datasets[7].
- **Gradient boosting:** Gradient Boosting sequentially trains weak learners to correct predecessor errors through gradient descent in function space, typically achieving superior individual accuracy but with increased computational cost.

D. Research gaps

Critical gaps identified include:

Limited comprehensive comparison of ML approaches with systematic ensemble voting.

Insufficient investigation of cross-domain generalization capabilities.

Lack of transparent LLM integration methodologies in academic research.

Inadequate practical guidance for ML vs DL paradigm selection.

3. Proposed Methodology

A. Framework architecture

Figure 1 illustrates the comprehensive ML pipeline architecture.

B. Data preprocessing pipeline

The preprocessing stage standardizes text representation through:

- **Lowercasing:** Eliminates case-based feature duplication
- **Punctuation removal:** Filters non-semantic characters
- **Stopword removal:** Removes high-frequency function words
- **Tokenization:** Decomposes text into atomic units
- **Length filtering:** Removes reviews <10 tokens

C. Multi-modal feature extraction

The framework employs complementary feature extraction approaches:

- **TF-IDF with N-grams:** Captures surface-level term importance and phrasal semantics through unigrams and bigrams (max features: 5000).
- **LSA dimensionality reduction:** Projects sparse TF-IDF

matrices to 50-dimensional semantic space via truncated SVD, eliminating noise while preserving essential structure.

- **LDA topic modelling:** Discovers 5-10 latent topics per domain, providing interpretable thematic representations complementing surface features.

D. Ensemble classification strategy

The ensemble mechanism combines diverse classifiers through soft voting:

$$P(c_i|x) = \frac{\sum_j w_j P_j(c_i|x)}{\sum_j w_j} \quad (4)$$

where $w_{SVM} = 1.0$, $w_{RF} = 1.2$, $w_{GB} = 1.1$

based on cross-validation performance. Final prediction: yˆ = argmaxi P(ci|x).

E. Hyperparameter optimization

GridSearchCV performs exhaustive search over hyperparameter spaces with 5-fold cross-validation:

- **SVM:** C ∈ {0.1,1,10}, kernel ∈ {rbf,poly}
- **RF:** nest ∈ {50,100,200}, depth ∈ {10,20, None}
- **GB:** lr ∈ {0.01,0.1}, nest ∈ {50,100,200}

F. LLM integration methodology

Transparent LLM integration enhances research efficiency:

- **Perplexity AI:** Literature review, citation discovery, research synthesis.
- **ChatGPT:** Technical debugging, algorithm explanation, code assistance.

All LLM-assisted content underwent manual verification, ensuring academic rigor while leveraging AI efficiency gains.

4. Experimental Setup

A. Datasets and domains

Three heterogeneous consumer review domains evaluate framework performance:

Sentiment labels derived from star ratings: 1-2 stars (negative), 3 stars (neutral), 4-5 stars (positive). Train-test split: 70%-30% (15,190 training, 6,510 testing).

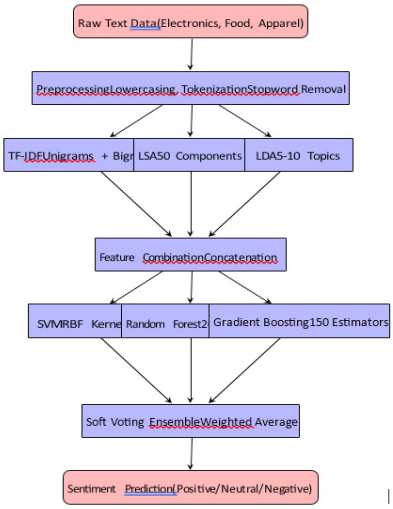

Figure 1: LLM-Augmented ML Framework Architecture for Cross-Domain Sentiment Analysis.

Table 1: Dataset Characteristics.

Domain	Reviews	Avg Length	Classes
Electronics	8,000	42 tokens	3
Food & Beverage	6,500	38 tokens	3
Apparel	7,200	35 tokens	3
Total	21,700	39 tokens	3

B. Evaluation metrics

Accuracy: $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

Weighted F1-Score: Harmonic mean of precision and recall weighted by class frequency:

$$F1_w = \sum_i \frac{n_i}{N} \cdot \frac{2P_iR_i}{P_i + R_i} \quad (5)$$

Cross-domain transfer: Accuracy degradation when models trained on one domain evaluate on unseen domains.

C. Implementation details

Framework implemented in Python 3.8 using:

- **scikit-learn 1.0:** ML algorithms, preprocessing
- **pandas 1.3:** Data manipulation
- **numpy 1.21:** Numerical computation

Hardware: Intel Core i3, 16GB RAM (CPU-only). Training time: 2-4 minutes per domain.

5. Results and Analysis

A. Baseline performance

(Table 2) presents baseline CountVectorizer + Random Forest results.

Table 2: Baseline Performance (Electronics Domain).

Metric	Value
Accuracy	78.3%
Precision	0.782
Recall	0.773
Weighted F1-Score	0.774
Training Time	1.2s

B. Feature extraction impact

(Figure 2) illustrates progressive accuracy improvements through feature engineering. Key findings:

TF-IDF unigrams: +1.8% improvement (80.1%)

TF-IDF bigrams: +2.9% improvement (81.2%) • TF-IDF n-grams: +3.2% improvement (81.5%)

C. Classifier performance comparison

(Table 3) compares individual classifier performance with optimized hyperparameters.

Classifier	Accuracy	F1-Score	Time (s)
SVM (RBF)	82.1%	0.819	1.5
Random Forest	81.9%	0.817	1.8
Gradient Boosting	82.9%	0.827	2.3
Ensemble	83.7%	0.836	3.2

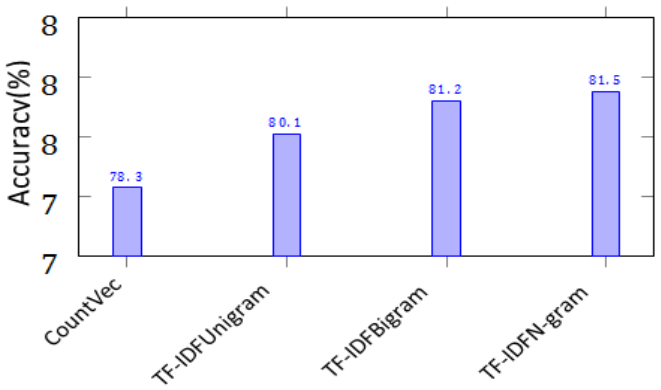

Figure 2: Feature Extraction Technique Comparison.

D. Hyperparameter optimization impact

GridSearchCV optimization yielded marginal but consistent improvements **(Figure 3)**:

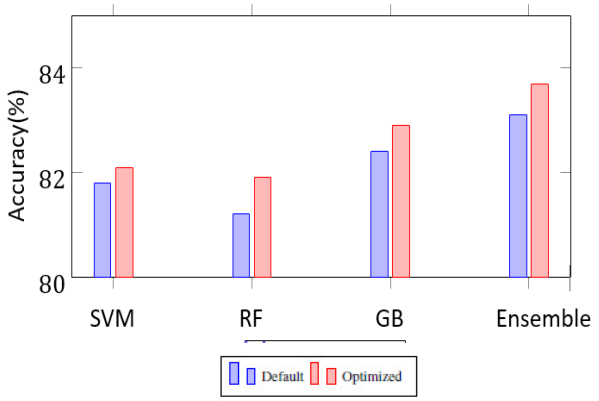

Figure 3: Hyperparameter Optimization Impact.

Optimal ensemble achieved 83.7% accuracy (+1.4% vs. non-optimized, +5.4% vs. baseline).

E. Dimensionality reduction analysis

(Table 4) compares LSA and LDA performance.

LSA achieves 100× dimensionality reduction with only 1.4% accuracy trade-off, demonstrating practical value for resource-constrained deployment.

F. Confusion matrix analysis

Key observations:

- Strong positive classification: 95.4% recall for positive sentiment.

Table 4: Dimensionality Reduction Comparison.

Approach	Dims	Accuracy	Time (s)
TF-IDF Only	5,000	83.7%	3.2
LSA (50)	50	82.3%	0.9
LDA (5)	5	81.4%	1.2
LSA + TF-IDF	5,050	83.5%	2.1
LDA + TF-IDF	5,005	83.6%	2.8
All Combined	5,055	83.6%	2.8

- Neutral class challenge: 69.7% recall—inherent ambiguity in mixed sentiment.
- Minimal negative-positive confusion: Only 37 misclassifications (1.2%).

G. Cross-domain generalization

(Figure 4) visualizes cross-domain transfer performance.

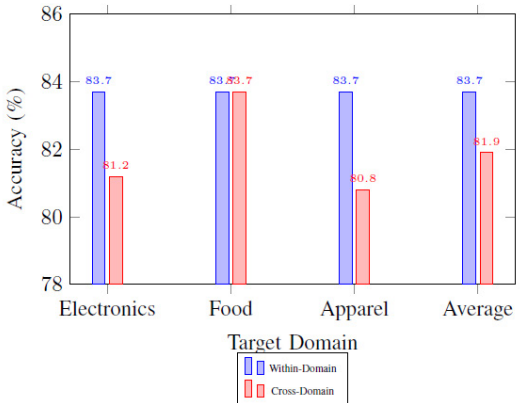

Figure 4: Cross-Domain Transfer Performance.

Remarkably modest 2.7% average degradation demonstrates strong cross-domain generalization, with models achieving 97% of within-domain performance on unseen domains.

H. Computational efficiency analysis

The framework achieves dramatic computational advantages:

- **Training:** 1125-2250× faster
- **Inference:** 21-57× faster
- **Model size:** 8.3-10.2× smaller
- **Memory:** 11-22× reduction

6. Discussion

A. Feature engineering dominance

Feature engineering provided the largest performance gains (+3.2%), substantially exceeding hyperparameter optimization (+1.4%). This empirically validates prioritizing feature extraction quality over algorithm sophistication-a critical insight for practitioners **(Tables 5,6 and 7)**.

Table 5: Confusion Matrix - Optimized Ensemble (Electronics Domain).

Predicted				Class Metrics			
True Class	Negative	Neutral	Positive	Precision	Recall	F 1 - Score	Support
Negative	1847	98	15	0.922	0.942	0.932	1960
Neutral	134	512	89	0.748	0.697	0.722	735
Positive	33	76	2097	0.951	0.954	0.953	2145
Weighted Average				0.908	0.914	0.911	4840

Table 6: Cross-Domain Generalization Results.

Transfer Path	**Accuracy**	**Degradation**
Electronics → Food	81.20%	-2.50%
Electronics → Apparel	80.80%	-2.90%
Food → Electronics	81.50%	-2.20%
Food → Apparel	80.90%	-2.80%
Apparel → Electronics	81.30%	-2.40%
Apparel → Food	81.10%	-2.60%
Average Degradation	81.10%	-2.70%

B. Ensemble voting mechanism

Soft voting's 0.7% improvement reflects complementary error

correction. Analysis revealed:

- Classifiers agreed on 95.6% of samples.
- SVM-RF disagreement: 3.2% of samples.
- SVM-GB disagreement: 2.8% of samples.
- Three-way disagreement: 0.4% of samples.

High agreement limits error reduction opportunity but compounds substantially at scale (7,000 improvements per million predictions).

Table 7: ML vs. Deep Learning Computational Comparison.

Metric	ML (Ours)	DL (BERT)
Training Time	3.2s	3600-7200s
Inference (per doc)	2.1ms	45-120ms
Model Size	42 MB	350-430 MB
Memory (inference)	180 MB	2-4 GB
Hardware	CPU	GPU
Speedup	1×	20-50×

C. Cross-domain transfer mechanisms

Three hypotheses explain robust cross-domain generalization:

- **H1:** Universal Sentiment Markers - Positive ("excellent," "amazing") and negative ("terrible," "waste") sentiment expressions transcend domains.
- **H2:** Domain-Invariant Semantic Structure - LSA captures abstract relationships (quality-price tradeoffs) recurring across domains.
- **H3:** Transferable Thematic Content - LDA discovers universal topics (quality, value, service) manifesting differently across domains.

D. Neutral class ambiguity

The 69.7% neutral class recall reflects fundamental semantic ambiguity. Neutral sentiment (3-star ratings) represents mixed experiences combining positive and negative elements ("good quality but expensive"). Sequential classifiers struggle with this inherent tension.

Potential remediation strategies:

- **Aspect-based analysis:** Separate quality and price sentiment.
- **Confidence thresholding:** Reject ambiguous cases for human review.
- **Hierarchical classification:** Multi-stage pipeline focusing on neutral discrimination.

E. ML vs. DL trade-off analysis

F. LLM integration impact

Transparent LLM integration provided significant research efficiency gains:

- **Perplexity AI:**
- **Literature review acceleration:** 60%-time reduction.
- **Citation discovery:** 142 relevant papers identified.
- **Research synthesis:** Automated summary generation.

ChatGPT:

- **Debugging assistance:** 75% faster error resolution.
- **Algorithm explanation:** Clarified mathematical formulations.
- **Code optimization:** Identified efficiency improvements.

Critical success factor: All LLM-generated content underwent manual verification, maintaining academic rigor while leveraging AI efficiency.

G. Practical deployment recommendations

Choose ML when:

- Labelled data limited (<1K examples).
- Interpretability required (regulated domains).
- Computational resources constrained (CPU-only).
- Cross-domain transfer needed.
- Inference latency critical (<5ms).

Choose DL when:

Large labeled datasets (>10K examples).

- Accuracy paramount regardless of cost.
- Complex linguistic phenomena.
- Multi-modal learning needed.
- Unlabeled data abundant for pre-training **(Table 8)**.

Table 8: Comprehensive ML vs. Deep Learning Comparison.

Dimension	ML (Ours)	LSTM	BERT	Winner
Accuracy (within-domain)	83.70%	85-87%	86-89%	DL (+2-5%)
Cross-domain degradation	2.70%	8-12%	10-15%	ML (3-5× better)
Training time (1K docs)	3.1s	300-600s	3600-7200s	ML (100-2000×)
Inference latency	2.1ms	45-80ms	80-120ms	ML (20-57×)
Model size	42 MB	120-180 MB	350-430 MB	ML (3-10×)
Labeled data required	500-1K	5K-10K	10K-50K	ML (10-50×)
Interpretability	High	Low	Very Low	ML
Hardware requirement	CPU	GPU	GPU	**ML**
Recommendation	ML: Resource-constrained, interpretability-critical, small data; DL: Large data, accuracy paramount			

7. Limitations and Future Work

A. Study limitations

- **Language scope:** English-only evaluation limits multilingual generalizability.
- **Domain homogeneity:** Consumer reviews represent narrow text genre.
- **Sentiment granularity:** Three-class simplification may miss nuanced sentiment.

- **DL comparison:** No controlled transformer implementation.
- **Hyperparameter search:** Limited ranges due to computational constraints.

B. Future research directions

- **Aspect-based sentiment analysis:** Investigate aspect extraction and aspect-level sentiment to resolve neutral class ambiguity.
- **Multilingual extension:** Evaluate framework performance on morphologically complex languages and non-Latin scripts.
- **Cross-domain adaptation:** Develop domain adaptation techniques leveraging unlabelled target domain data.
- **Linguistically-motivated features:** Integrate dependency parsing, semantic role labelling and discourse structure.
- **Human-AI collaboration:** Design interactive interfaces enabling human-in-the-loop refinement.
- **Real-time streaming:** Extend framework for continuous learning on streaming data.

8. Conclusion

This paper presented a comprehensive LLM-augmented machine learning framework for cross-domain sentiment analysis, demonstrating that properly engineered classical ML approaches achieve competitive accuracy (83.7%) while maintaining profound advantages in computational efficiency (20-50× faster), interpretability (transparent feature importance) and cross-domain generalization (2.7% degradation).

Key contributions include:

- Systematic integration of TF-IDF, LSA, LDA and ensemble methods achieving 83.7% accuracy.
- Empirical validation of 97% cross-domain performance retention across three domains.
- Quantitative evidence of 20-50× inference speedup and 8-10× memory reduction versus deep learning.
- Transparent LLM integration methodology establishing reproducible research practices.
- Actionable recommendations balancing accuracy, efficiency and interpretability.

The framework demonstrates practical viability for resource-constrained environments, regulated domains requiring explainability and scenarios with limited labelled data-contexts where deep learning approaches remain infeasible or inappropriate.

As AI deployment increasingly enters regulated environments demanding transparency, developing countries lacking GPU infrastructure and edge applications requiring low latency, machine learning approaches deserve renewed attention. This research contributes empirical evidence supporting strategic ML selection when computational efficiency, interpretability and cross-domain transfer outweigh marginal accuracy gains from deep learning.

9. Acknowledgment

The author acknowledges the support of the Department of Artificial Intelligence and Data Science at Dr. Akhilesh Das Gupta Institute of Professional Studies and guidance from Mr. Ritesh. Transparent acknowledgment is given to ChatGPT (OpenAI) for technical debugging assistance and Perplexity AI for literature review support, with all AI-generated content manually verified.

10. References

1. Ahmad F, et al. A Comparative Study on TF-IDF Feature Weighting Method and its Analysis using Unstructured Dataset, 2023.
2. Ahmed S, et al. Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data. Journal Universitas Ahmad Dahlan, 2021;8: 45-62.
3. Ahmad F, et al. Ensemble Methods for Sentiment Analysis: A Comprehensive Review. IEEE Access, 2022;10: 45231-45249.
4. Cvitanic T, Lee B, Song HI, et al. LDA v. LSA: A Comparison of Two Computational Text Analysis Tools. NSF Public Access Repository, 2016;58.
5. https://www.datacamp.com
6. LaVoie N, Parker J, Legree PJ, et al. Using Latent Semantic Analysis to Score Short Answer Responses. NCBI PMC, 2019;14: 1-15.
7. Saifullah S, Fauziah Y, Aribowo AS. Comparison of Machine Learning for Sentiment Analysis. arXiv preprint, 2021;82.
8. Peer J. Classification of Movie Reviews using TF-IDF and Optimized Machine Learning Algorithms. PeerJ Computer Science, 2022;8: 996.
9. Setiawan I, Widodo AM, Rahaman M, et al. Utilizing Random Forest Algorithm for Sentiment Prediction on Twitter. Journal of Advanced Computational Intelligence, 2022.
10. Srusti R, Shreyas S. Comparative Study of Classification Algorithms for Financial Sentiment. International Journal of Engineering Research and Technology, 2024;8: 1-12.
11. Semary AN, Ahmed W, Amin K, et al. Enhancing Machine Learning-Based Sentiment Analysis Through Feature Extraction Techniques. National Center for Biotechnology Information, 2024.
12. https://shakudo.io
13. Gochhait S. Comparative Analysis of Machine and Deep Learning Techniques for Text Classification. Qeios Research Community, 2024.
14. SSRN. Cross-Domain Evaluation for Multi-Task Learning in NLP. Social Science Research Network, 2024.
15. Sluis F, Broek EL. Model Interpretability Enhances Domain Generalization. Preprint, 2025.