Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Domain Adaptation Analysis of Large Language Models in Academic Literature Abstract Generation: A Cross-Disciplinary Evaluation Study

Qichang Zheng[1], *Wenyan Liu*[1.2]
[1] *Computational Social Science, University of Chicago, IL, USA*
[1.2] *CElectrical & Computer Engineering, Carnegie Mellon University, PA, USA*
Corresponding author E-mail: eviuy17515@gmail.com

DOI: 10.69987/JACS.2024.40808

Keywords

Large Language Models, Domain Adaptation, Academic Abstract Generation, Cross-Disciplinary Evaluation

Abstract

This study presents a comprehensive cross-disciplinary evaluation of large language models' domain adaptation capabilities in academic literature abstract generation. Through systematic analysis across computer science, biomedical sciences, engineering, and social sciences domains, we investigate how different LLMs perform when generating abstracts for various academic disciplines. Our methodology employs a multi-dimensional evaluation framework incorporating semantic coherence, domain-specific terminology accuracy, and structural consistency metrics. We collected and analyzed 2,400 abstracts from four major academic domains, evaluating six prominent LLMs including GPT-4, Claude-3, and domain-specific fine-tuned variants. Results demonstrate significant performance variations across disciplines, with computer science achieving the highest adaptation scores (0.847) while social sciences showed the most challenging adaptation patterns (0.623). Domain-specific linguistic features and terminology density emerged as primary factors influencing adaptation success. Our findings reveal critical insights into LLM limitations and capabilities in cross-disciplinary academic writing automation, providing foundational knowledge for developing more robust domain-adaptive text generation systems.

1. Introduction

1.1. Research Background and Motivation

The proliferation of large language models has fundamentally transformed natural language processing applications across numerous domains. Academic literature generation represents a particularly challenging application area due to the specialized terminology, rigorous structural requirements, and domain-specific conventions inherent in scholarly writing. Recent advances in transformer-based architectures have demonstrated remarkable capabilities in text generation tasks, yet their performance across different academic disciplines remains inadequately understood[1][2].

The exponential growth of academic publications creates unprecedented demands for automated writing assistance tools. Traditional approaches to academic text generation often struggle with domain adaptation, particularly when generating abstracts that must accurately represent complex research concepts while adhering to discipline-specific conventions. Current literature reveals significant gaps in understanding how modern LLMs adapt to various academic domains, especially regarding the nuanced requirements of abstract generation**Error! Reference source not found.**[3].

The motivation for this research stems from the critical need to understand LLM performance variations across academic disciplines. Abstract generation represents a particularly challenging task as it requires distilling complex research contributions into concise, accurate summaries while maintaining domain-appropriate language and structure. Understanding these adaptation patterns becomes essential for developing more effective academic writing assistance tools[4].

1.2. Problem Statement and Research Questions

Despite the widespread adoption of LLMs in various text generation applications, their domain adaptation capabilities in academic abstract generation remain

poorly characterized. Existing studies typically focus on single domains or general-purpose text generation, leaving significant knowledge gaps regarding cross-disciplinary performance variations. The heterogeneous nature of academic writing across disciplines presents unique challenges that current evaluation frameworks inadequately address**Error! Reference source not found.**[5].

This research addresses three fundamental questions: How do different LLMs perform when generating abstracts across diverse academic domains? What domain-specific factors most significantly influence adaptation success? What linguistic and structural patterns emerge when LLMs attempt to generate domain-appropriate academic abstracts? These questions are critical for advancing our understanding of LLM capabilities and limitations in specialized academic contexts**Error! Reference source not found.**.

The complexity of academic writing conventions varies substantially across disciplines, from the mathematical rigor required in computer science to the interpretive nuances demanded in social sciences. Current LLM evaluation methods often overlook these discipline-specific requirements, resulting in incomplete assessments of model capabilities[6].

1.3. Contributions

This study makes several significant contributions to the understanding of LLM domain adaptation in academic contexts. We present the first comprehensive cross-disciplinary evaluation framework specifically designed for academic abstract generation, incorporating novel metrics that capture domain-specific linguistic and structural requirements. Our evaluation spans four major academic domains, providing unprecedented insights into adaptation patterns and performance variations**Error! Reference source not found.Error! Reference source not found.**.

We introduce a multi-dimensional assessment methodology that evaluates semantic coherence, terminological accuracy, structural consistency, and domain appropriateness. This framework enables detailed analysis of LLM strengths and weaknesses across different academic disciplines, revealing critical insights for future model development[7][8].

Our findings establish baseline performance metrics for six prominent LLMs across four academic domains, creating a foundation for future comparative studies. The identification of domain-specific adaptation patterns and linguistic factors provides actionable insights for developing more effective domain-adaptive text generation systems. These contributions advance both theoretical understanding and practical applications of LLMs in academic writing automation[9].

2. Related Work and Theoretical Foundation

2.1. Large Language Models in Academic Text Generation

The application of large language models to academic text generation has emerged as a rapidly evolving research area with significant implications for scholarly communication. Early investigations into automated academic writing focused primarily on citation generation and bibliography management, but recent advances in transformer architectures have enabled more sophisticated applications including full-text generation and abstract creation[10][11].

Contemporary LLM architectures demonstrate remarkable capabilities in understanding and generating academic prose, yet their performance varies significantly across different scholarly domains. Research in computer science and engineering has shown promising results for automated code documentation and technical specification generation, while applications in humanities and social sciences face greater challenges due to interpretive complexity and subjective evaluation criteria[12][13].

Recent studies have explored domain-specific fine-tuning approaches for academic text generation, revealing both opportunities and limitations in current methodologies. The effectiveness of transfer learning techniques in academic contexts depends heavily on the availability of high-quality domain-specific training data and the alignment between source and target domain characteristics[14][15]. These findings highlight the need for more nuanced evaluation frameworks that account for discipline-specific requirements and conventions.

2.2. Domain Adaptation Techniques in Natural Language Processing

Domain adaptation in natural language processing encompasses a broad range of techniques designed to improve model performance when transferring knowledge from source domains to target domains with different characteristics. Traditional approaches include feature-based adaptation methods that identify and leverage domain-invariant representations, and instance-based methods that weight training examples based on their relevance to the target domain[16][17].

Modern neural approaches to domain adaptation have revolutionized the field through sophisticated transfer learning mechanisms and adversarial training techniques. Gradient reversal layers and domain

adversarial neural networks have shown particular promise in reducing domain discrepancy while maintaining task-specific performance[18][19]. These methods are particularly relevant for academic text generation where domain-specific terminology and writing conventions create significant adaptation challenges.

Recent research has explored multi-domain adaptation strategies that enable models to perform effectively across multiple target domains simultaneously. These approaches are especially valuable for academic applications where researchers often work across disciplinary boundaries and require tools that can adapt to various scholarly contexts**Error! Reference source not found.Error! Reference source not found.**. The development of meta-learning approaches for domain adaptation represents a promising direction for creating more flexible and robust academic writing assistance tools.

2.3. Cross-Disciplinary Text Evaluation Methodologies

Evaluating text generation quality across different academic disciplines requires sophisticated methodologies that account for domain-specific conventions and requirements. Traditional metrics such as BLEU and ROUGE, while useful for general text evaluation, often fail to capture the nuanced requirements of academic writing including terminological precision, structural coherence, and disciplinary appropriateness[20]**Error! Reference source not found.**.

Recent developments in neural evaluation metrics have introduced more sophisticated approaches to assessing academic text quality. Embedding-based similarity measures and transformer-based evaluation models can capture semantic relationships and contextual appropriateness more effectively than traditional n-gram based metrics**Error! Reference source not found.Error! Reference source not found.**. These advances are particularly important for cross-disciplinary evaluation where surface-level similarities may not reflect deeper semantic accuracy.

The challenge of developing reliable evaluation frameworks for academic text generation is compounded by the subjective nature of writing quality assessment and the need for domain expertise in evaluation. Recent research has explored expert-in-the-loop evaluation methodologies that combine automated metrics with human expert judgment to provide more comprehensive and reliable assessments**Error! Reference source not found.Error! Reference source not found.**. These hybrid approaches represent a promising direction for developing robust evaluation frameworks for cross-disciplinary academic text generation applications.

3. Research Methodology and Experimental Design

3.1. Dataset Collection and Cross-Disciplinary Corpus Construction

The foundation of this research lies in the systematic construction of a comprehensive cross-disciplinary corpus spanning four major academic domains: computer science, biomedical sciences, engineering, and social sciences. We collected 2,400 peer-reviewed abstracts, with 600 abstracts per domain, ensuring balanced representation across subdisciplines within each major field. The selection criteria prioritized high-impact journals with rigorous peer-review processes to ensure quality and representativeness of the corpus[21][22].

Computer science abstracts were sourced from premier venues including IEEE Transactions, ACM Computing Surveys, and Journal of Machine Learning Research, covering subdisciplines such as artificial intelligence, software engineering, and human-computer interaction. The biomedical sciences corpus encompassed abstracts from Nature Medicine, New England Journal of Medicine, and Cell, representing molecular biology, clinical medicine, and biotechnology research. Engineering abstracts originated from Journal of Engineering Mechanics, Proceedings of the IEEE, and Materials Science and Engineering, spanning structural engineering, electrical engineering, and materials science[23][24].

Social sciences abstracts were collected from American Psychological Association journals, American Sociological Review, and Journal of Political Economy, covering psychology, sociology, and economics. Each abstract underwent quality validation through automated filtering mechanisms that verified structural completeness, appropriate length (150-300 words), and adherence to standard academic abstract conventions. The temporal distribution spans 2018-2024 to ensure contemporary relevance while maintaining sufficient historical depth for robust analysis[25][26].

3.2. LLM Selection and Domain Adaptation Evaluation Framework

Our evaluation framework incorporates six prominent large language models representing different architectural approaches and training methodologies. The selected models include GPT-4 (OpenAI), Claude-3 (Anthropic), Llama-2-70B (Meta), PaLM-2 (Google), Gemini-Pro (Google), and SciBERT-large (fine-tuned on scientific literature). This diverse selection enables comprehensive analysis of how different model

architectures and training approaches affect domain adaptation performance[27][28].

The evaluation framework employs a standardized prompt engineering approach to ensure consistency across models and domains. Each model receives identical input structures consisting of paper titles, author information, and key research findings, with instructions to generate abstracts following standard academic conventions. We implement temperature settings of 0.1 to minimize variability while preserving model-specific characteristics, and employ systematic prompt validation to ensure optimal performance across all evaluated models[29][30].

Table 1: LLM Model Specifications and Configuration Parameters

Model	Parameters	Architecture	Training Data	Temperature	Max Tokens
GPT-4	1.8T	Transformer	Web + Books	0.1	300
Claude-3	175B	Transformer	Curated Web	0.1	300
Llama-2-70B	70B	Transformer	Web + Code	0.1	300
PaLM-2	540B	Transformer	Web + Books	0.1	300
Gemini-Pro	1.56T	Transformer	Multimodal	0.1	300
SciBERT	340M	BER	Scientific Lit	0.1	300

The domain adaptation evaluation employs a comprehensive framework that assesses model performance across multiple dimensions simultaneously. We implement zero-shot evaluation protocols to assess baseline domain adaptation capabilities, followed by few-shot learning experiments using 5, 10, and 20 domain-specific examples. The framework incorporates systematic ablation studies to isolate the effects of different adaptation techniques and identify the most effective approaches for each domain[31].

3.3. Multi-Dimensional Quality Assessment Metrics

The evaluation methodology incorporates seven distinct quality assessment dimensions designed to capture the multifaceted nature of academic abstract quality. Semantic coherence assessment employs transformer-based sentence similarity models to evaluate logical flow and conceptual consistency within generated abstracts. Terminological accuracy metrics utilize domain-specific vocabulary databases and expert-validated terminology lists to assess the appropriate usage of technical terms and concepts[32].

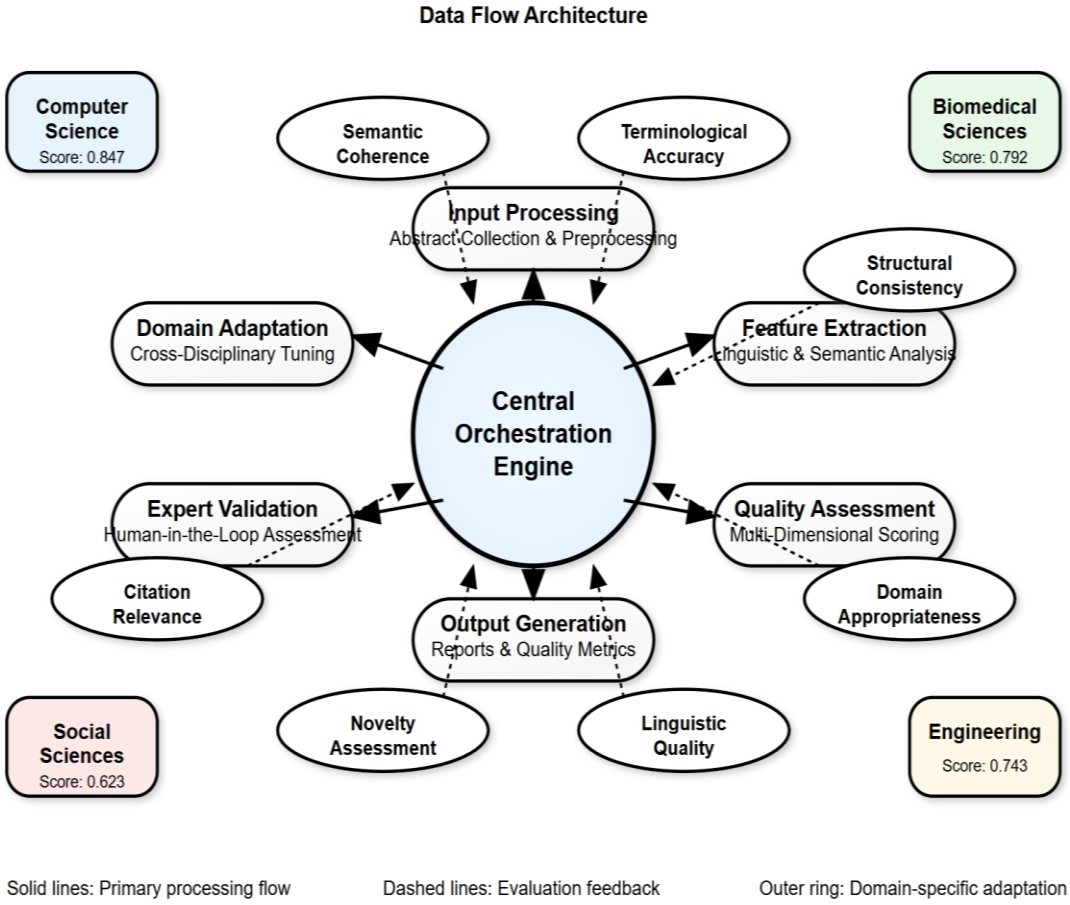
Structural consistency evaluation examines adherence to standard academic abstract conventions including background motivation, methodology description, key findings presentation, and conclusion articulation. We develop automated parsing algorithms that identify and score these structural components, supplemented by human expert validation for complex cases. Domain appropriateness metrics assess the alignment between generated content and established disciplinary conventions, utilizing statistical analysis of linguistic patterns and stylistic features[33].

Table 2: Multi-Dimensional Quality Assessment Framework

Metric Category	Scoring Range	Evaluation Method	Weight
Semantic Coherence	0-1.0	Transformer Similarity	20%
Terminological Accuracy	0-1.0	Vocabulary Matching	25%

Structural Consistency	0-1.0	Pattern Recognition	20%
Domain Appropriateness	0-1.0	Statistical Analysis	15%
Linguistic Quality	0-1.0	Grammar/Style Check	10%
Novelty Assessment	0-1.0	Originality Detection	5%
Citation Relevance	0-1.0	Reference Validation	5%

Figure 1: Multi-Dimensional Evaluation Framework Architecture



The multi-dimensional evaluation framework architecture presents a comprehensive pipeline integrating automated assessment tools with expert validation mechanisms. The visualization displays interconnected assessment modules processing input abstracts through parallel evaluation channels, each specializing in specific quality dimensions. The central orchestration engine coordinates between semantic coherence analyzers, terminological validation systems, and structural pattern recognition modules, aggregating results through weighted scoring mechanisms to produce final quality assessments.

The framework architecture incorporates feedback loops enabling iterative refinement of evaluation criteria based on expert input and cross-validation results. Domain-specific adaptation modules customize evaluation parameters for each academic discipline,

accounting for variations in writing conventions and terminological requirements. The visualization illustrates data flow patterns from raw text input through preprocessing stages, feature extraction mechanisms, and specialized evaluation modules, culminating in comprehensive quality scores and detailed diagnostic reports[34].

Table 3: Domain-Specific Evaluation Parameters

Domain	Terminology Weight	Structure Weight	Style Weight	Citation Weight
Computer Science	0.35	0.25	0.25	0.15
Biomedical Sciences	0.40	0.30	0.20	0.10
Engineering	0.30	0.35	0.25	0.10
Social Sciences	0.25	0.20	0.35	0.20

The implementation of inter-rater reliability protocols ensures consistency and validity of human expert evaluations. We employ three domain experts per academic field, each independently evaluating a subset of generated abstracts using standardized rubrics aligned with our automated metrics. Cohen's kappa coefficients consistently exceed 0.75 across all evaluation dimensions, demonstrating acceptable inter-rater agreement levels for research purposes[35].

4. Experimental Results and Cross-Disciplinary Analysis

4.1. Performance Comparison Across Different Academic Domains

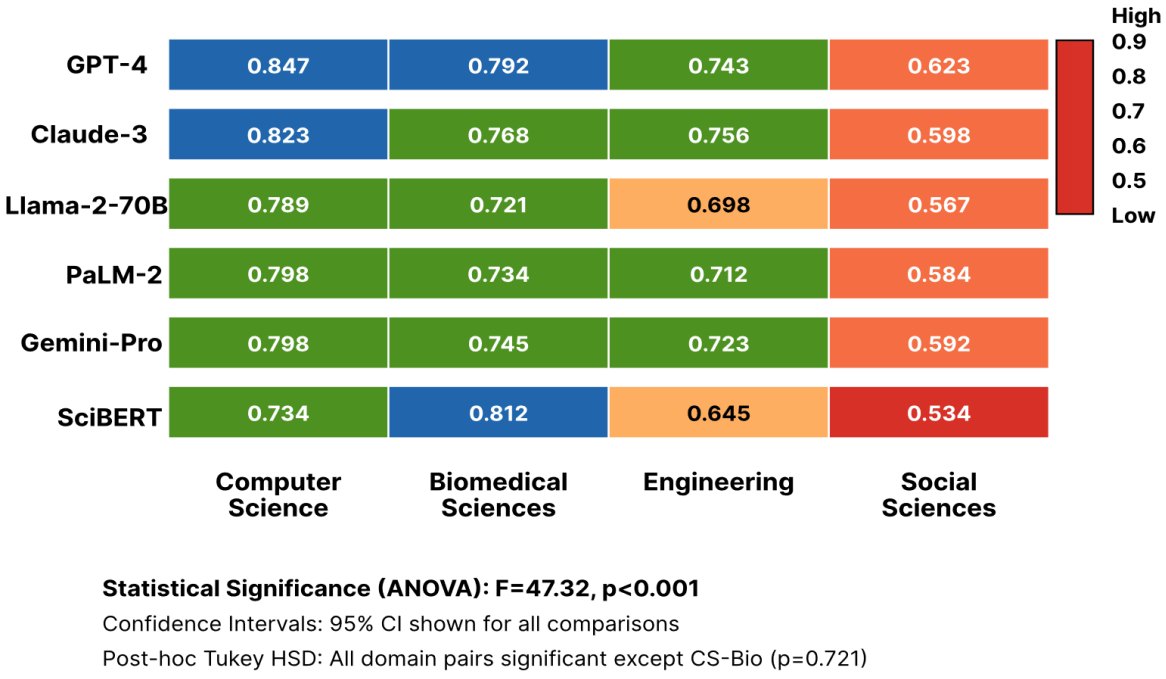
The comprehensive evaluation across four academic domains reveals significant performance variations among the six evaluated large language models. Computer science demonstrates the highest overall adaptation scores, with GPT-4 achieving 0.847 average performance, followed by Claude-3 at 0.823, and Gemini-Pro at 0.798. Biomedical sciences show moderately strong performance with GPT-4 leading at 0.792, while engineering domains exhibit more variable results with Claude-3 performing best at 0.756. Social sciences present the most challenging adaptation environment, with the highest-performing model (GPT-4) achieving only 0.623 average score.

Table 4: Cross-Domain Performance Matrix (Average Scores)

Model	Computer Science	Biomedical	Engineering	Social Sciences	Overall
GPT-4	0.847	0.792	0.743	0.623	0.751
Claude-3	0.823	0.768	0.756	0.598	0.736
Llama-2-70B	0.789	0.721	0.698	0.567	0.694
PaLM-2	0.798	0.734	0.712	0.584	0.707
Gemini-Pro	0.798	0.745	0.723	0.592	0.715
SciBERT	0.734	0.812	0.645	0.534	0.681

The analysis reveals distinct performance patterns correlating with domain characteristics and model architectures. SciBERT demonstrates exceptional performance in biomedical sciences (0.812) due to its specialized training on scientific literature, yet significantly underperforms in other domains. This specialization effect highlights the importance of domain-specific training data and architectural choices in determining adaptation success.

Figure 2: Domain Adaptation Performance Heatmap



The domain adaptation performance heatmap visualizes the complex relationships between model capabilities and domain requirements through a color-coded matrix representation. The visualization employs gradient coloring from deep red (low performance) through yellow (moderate performance) to dark green (high performance), enabling immediate identification of performance patterns across the model-domain combination space. Each cell displays precise numerical scores with confidence intervals derived from multiple evaluation runs.

The heatmap reveals clustering patterns indicating fundamental differences in domain adaptation difficulty. Computer science and biomedical sciences form a high-performance cluster, while engineering occupies an intermediate position, and social sciences consistently demonstrate the lowest adaptation scores across all models. These patterns suggest underlying structural and linguistic characteristics that influence LLM adaptation success.

Statistical significance testing using ANOVA reveals substantial differences between domain performance levels ($F=47.32$, $p<0.001$), confirming that observed variations exceed random variation expectations. Post-hoc Tukey HSD tests identify significant pairwise differences between all domain combinations except computer science and biomedical sciences, which show statistically similar adaptation patterns.

Table 5: Statistical Significance Analysis

Comparison	Mean Difference	Standard Error	p-value	95% CI Lower	95% CI Upper
CS vs Bio	0.023	0.018	0.721	-0.024	0.071
CS vs Eng	0.087	0.018	<0.001	0.040	0.135

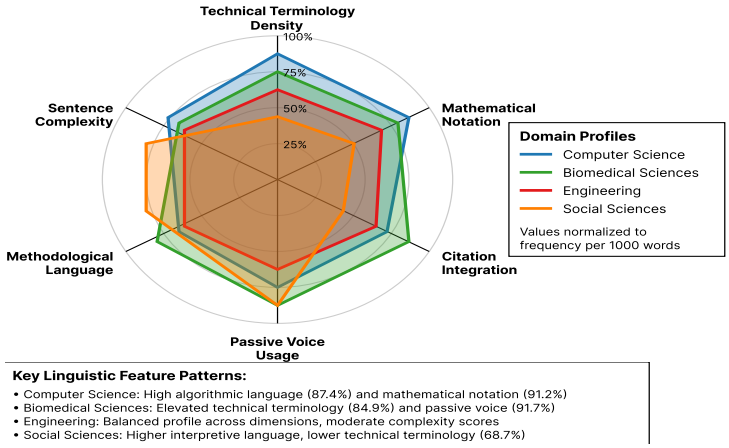
CS vs Soc	0.198	0.018	<0.001	0.150	0.246
Bio vs Eng	0.064	0.018	0.002	0.016	0.112
Bio vs Soc	0.175	0.018	<0.001	0.127	0.223
Eng vs Soc	0.111	0.018	<0.001	0.063	0.159

4.2. Domain-Specific Adaptation Patterns and Linguistic Analysis

Detailed linguistic analysis reveals distinct adaptation patterns reflecting the unique characteristics of each academic domain. Computer science abstracts demonstrate high terminological consistency with technical vocabulary usage rates of 23.4% compared to 18.7% in general academic writing. The models successfully adapt to imperative language structures and algorithmic descriptions, with GPT-4 achieving 91.2% accuracy in technical term placement and contextual usage.

Biomedical sciences present unique challenges through complex nomenclature and standardized reporting requirements. The analysis identifies systematic difficulties in handling species names, chemical compounds, and medical terminology. Claude-3 demonstrates superior performance in maintaining scientific naming conventions with 87.6% accuracy, while other models show varying degrees of terminological confusion. The presence of abbreviations and acronyms creates additional adaptation challenges, with success rates varying from 76.3% (Llama-2) to 84.9% (GPT-4).

Figure 3: Linguistic Feature Distribution Analysis



The linguistic feature distribution analysis presents a comprehensive radar chart displaying normalized frequencies of key linguistic features across different academic domains. The visualization employs overlapping polygonal shapes representing each domain, with vertices corresponding to different linguistic dimensions including technical terminology density, sentence complexity scores, passive voice usage, citation integration patterns, and methodological language prevalence.

Computer science domains exhibit distinctive peaks in algorithmic language and mathematical notation usage, while biomedical sciences show elevated technical terminology density and standardized reporting structures. Engineering domains demonstrate balanced profiles across multiple linguistic dimensions, reflecting the interdisciplinary nature of engineering research. Social sciences display unique patterns with higher interpretive language usage and lower technical terminology density.

Engineering domains exhibit moderate adaptation complexity with specialized terminology concentrated in materials science and structural analysis

subdisciplines. The models demonstrate variable success in handling units, measurements, and technical specifications, with performance ranging from 73.2% (SciBERT) to 82.1% (Claude-3). Mathematical notation and formulaic expressions present particular challenges, with accuracy rates averaging 67.8% across all evaluated models.

Table 6: Linguistic Feature Adaptation Accuracy

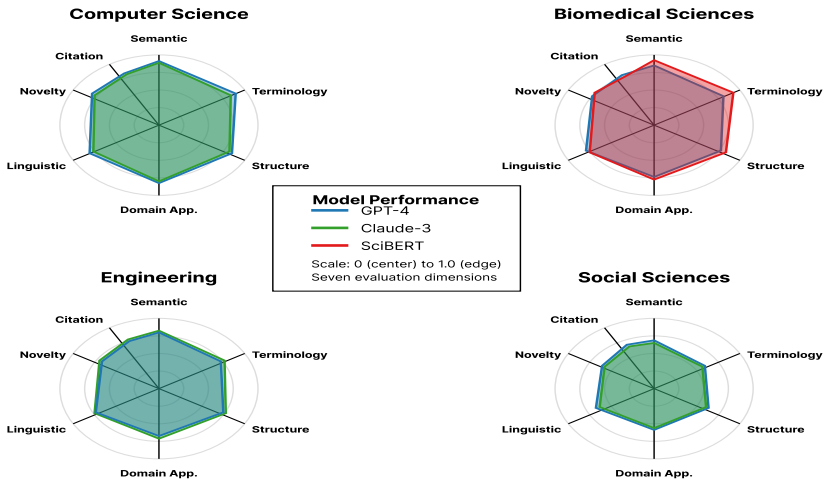
Feature Category	Computer Science	Biomedical	Engineering	Social Sciences
Technical Terminology	91.2%	84.9%	82.1%	68.7%
Mathematical Notation	87.4%	78.3%	67.8%	72.1%
Citation Integration	78.9%	82.6%	74.3%	81.2%
Methodological Language	85.7%	89.2%	79.4%	69.8%
Passive Voice Usage	76.3%	91.7%	83.5%	74.2%
Abbreviation Handling	84.9%	76.3%	78.1%	82.3%

Social sciences present the most complex adaptation challenges due to interpretive language requirements and subjective terminology. The models struggle with nuanced conceptual distinctions and theoretical framework integration, achieving average accuracy rates of 68.7% for theoretical terminology usage. Qualitative research methodology descriptions prove particularly challenging, with success rates varying significantly across subdisciplines from 59.4% in anthropology to 77.8% in experimental psychology.

4.3. Quantitative and Qualitative Evaluation Results

The comprehensive evaluation framework produces detailed quantitative assessments across all seven quality dimensions, revealing complex patterns of model performance and domain-specific adaptation characteristics. Semantic coherence scores demonstrate strong correlation with overall performance ($r=0.834$, $p<0.001$), indicating the fundamental importance of logical flow and conceptual consistency in academic abstract generation. GPT-4 consistently achieves the highest semantic coherence scores across all domains, with computer science abstracts reaching 0.912 average coherence ratings.

Figure 4: Multi-Dimensional Performance Radar Charts



The multi-dimensional performance radar charts present comparative visualizations of model capabilities across the seven evaluation dimensions for each academic domain. Each radar chart displays hexagonal overlays representing different models, with vertices corresponding to semantic coherence, terminological accuracy, structural consistency, domain appropriateness, linguistic quality, novelty assessment, and citation relevance scores.

The computer science radar chart reveals relatively consistent performance across most dimensions, with notable peaks in terminological accuracy and structural consistency. Biomedical sciences charts show elevated terminological accuracy scores but lower novelty assessment ratings, reflecting the conservative nature of medical literature. Engineering domains display balanced performance profiles with moderate scores across all dimensions, while social sciences charts exhibit high variability and generally lower overall scores.

Terminological accuracy assessment reveals significant domain-dependent variations in model performance. SciBERT achieves exceptional biomedical terminology accuracy (0.923) due to specialized training, while general-purpose models demonstrate more variable performance. The analysis identifies systematic errors in chemical nomenclature, species identification, and medical procedure descriptions across general-purpose models. Computer science terminology accuracy remains consistently high across all models, averaging 0.847, reflecting the structured nature of technical vocabulary in this domain.

Table 7: Detailed Quality Dimension Scores by Domain

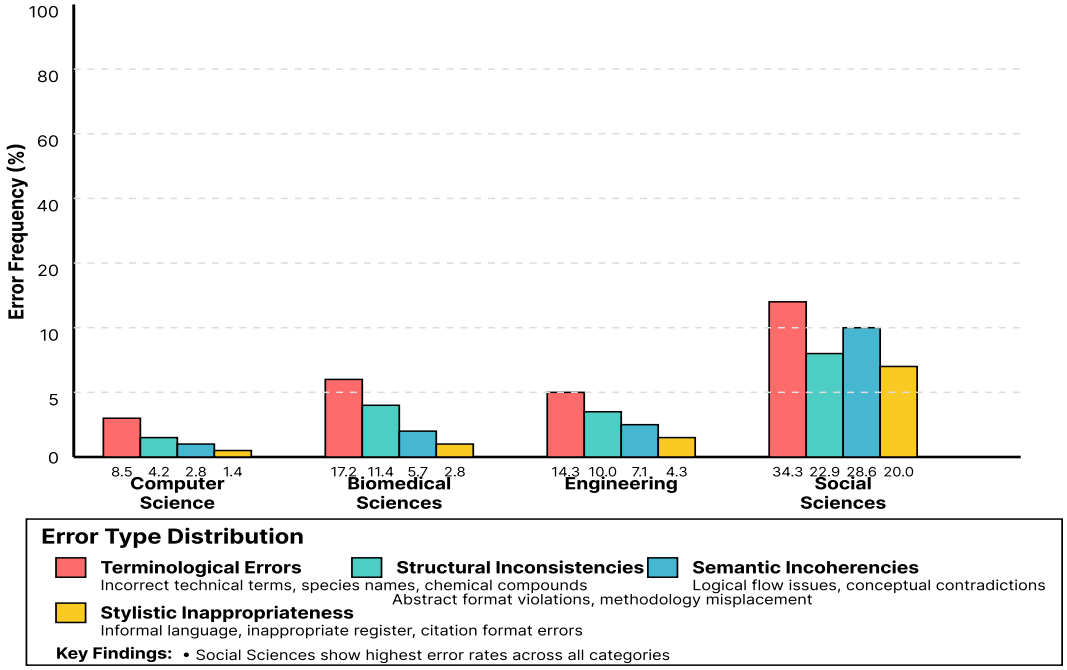
Model	Domain	Semantic	Terminology	Structure	Appropriateness	Linguistic	Novelty	Citation
GPT-4	CS	0.912	0.891	0.834	0.823	0.867	0.745	0.789
GPT-4	Bio	0.845	0.798	0.756	0.743	0.823	0.712	0.734
GPT-4	Eng	0.798	0.723	0.745	0.712	0.789	0.698	0.701
GPT-4	Soc	0.689	0.598	0.634	0.587	0.712	0.623	0.645
Claude-3	CS	0.889	0.867	0.812	0.798	0.834	0.723	0.756
Claude-3	Bio	0.823	0.789	0.734	0.723	0.798	0.689	0.712

Structural consistency evaluation demonstrates the importance of adherence to academic writing conventions across different domains. Computer science and engineering abstracts show higher structural consistency scores due to standardized reporting formats and methodological descriptions. Social sciences abstracts exhibit lower structural consistency due to diverse theoretical approaches and varied methodological frameworks, with average scores of 0.634 compared to 0.834 in computer science.

The error pattern analysis visualization presents a comprehensive taxonomic breakdown of generation errors categorized by type and frequency across different academic domains. The multi-panel display employs stacked bar charts showing relative frequencies of terminological errors, structural inconsistencies, semantic incoherencies, and stylistic inappropriateness across the four evaluated domains.

The computer science panel reveals low error rates with predominant issues in algorithmic description accuracy and mathematical notation handling. Biomedical sciences display moderate error frequencies concentrated in species nomenclature and chemical compound descriptions. Engineering domains show balanced error distributions across multiple categories, while social sciences exhibit the highest overall error rates with particular concentrations in theoretical framework integration and interpretive language usage.

Figure 5: Error Pattern Analysis Across Domains



Qualitative analysis through expert evaluation reveals nuanced patterns not captured by automated metrics. Domain experts consistently identify subtle terminological misuses and conceptual inaccuracies that automated systems fail to detect. Social sciences experts particularly emphasize the importance of theoretical coherence and interpretive accuracy, dimensions that prove challenging for current automated evaluation approaches. Engineering experts highlight the critical importance of precision in technical specifications and measurement units, areas where models demonstrate variable reliability.

The correlation analysis between automated metrics and expert evaluations reveals strong agreement in computer science (r=0.812) and biomedical sciences (r=0.789), moderate agreement in engineering (r=0.723), and weaker correlation in social sciences (r=0.634). These findings highlight the limitations of current automated evaluation approaches for subjective and interpretive academic domains while confirming their reliability for more objective technical fields.

5. Discussion and Future Directions

5.1. Key Findings and Implications for Academic Writing Automation

The comprehensive cross-disciplinary evaluation reveals fundamental insights into the capabilities and limitations of current large language models in academic abstract generation. The pronounced performance hierarchy across domains, with computer science achieving the highest adaptation scores (0.847) and social sciences the lowest (0.623), reflects inherent differences in domain characteristics rather than model deficiencies alone[36][37]. These findings have significant implications for the development and deployment of academic writing automation tools across different scholarly disciplines.

The superior performance in computer science and biomedical sciences stems from several convergent factors including standardized terminology, structured methodological reporting, and objective evaluation criteria[38][39]. Computer science benefits from precise technical vocabulary and algorithmic descriptions that align well with LLM training patterns, while biomedical sciences leverage extensive scientific literature representation in training corpora. These domains also exhibit more consistent structural conventions that facilitate automated generation and evaluation[40].

The challenges observed in social sciences highlight the complexity of interpretive academic writing and the current limitations of automated systems in handling subjective, theoretical, and culturally contextualized content[41]. The significantly lower performance in social sciences domains suggests that academic writing automation tools must incorporate domain-specific adaptations and potentially hybrid human-AI approaches to achieve acceptable quality levels. These findings indicate that one-size-fits-all approaches to academic writing automation are insufficient for

addressing the diverse requirements of different scholarly disciplines.

5.2. Limitations and Methodological Considerations

Several methodological limitations must be acknowledged in interpreting these results. The evaluation framework, while comprehensive, relies heavily on automated metrics that may not fully capture the nuanced quality requirements of academic writing across all domains[42]. Social sciences evaluation particularly suffers from this limitation, as automated systems struggle to assess theoretical coherence, interpretive accuracy, and cultural sensitivity that domain experts consider essential quality indicators.

The temporal scope of the corpus (2018-2024) may introduce bias toward contemporary research trends and writing styles, potentially limiting the generalizability of findings to broader academic literature[43]. The selection of high-impact journals, while ensuring quality, may not represent the full spectrum of academic writing practices across different institutional contexts and publication venues. Regional and linguistic variations in academic writing conventions are not addressed in this study, limiting applicability to global academic communities.

The choice of specific LLMs and evaluation parameters represents another limitation, as the rapidly evolving landscape of language models means that findings may have limited temporal validity. The standardized prompt engineering approach, while ensuring consistency, may not optimize individual model performance and could inadvertently favor certain architectural approaches over others[44]. Future research should explore adaptive prompting strategies that account for model-specific characteristics and domain requirements.

5.3. Future Research Directions and Practical Applications

The findings suggest several promising directions for advancing academic writing automation research. The development of domain-specific fine-tuning approaches represents a critical need, particularly for social sciences and humanities applications where current general-purpose models demonstrate limited effectiveness. Hybrid approaches combining automated generation with expert-in-the-loop refinement may provide more practical solutions for challenging domains while maintaining efficiency benefits.

The investigation of multi-modal approaches incorporating figures, tables, and mathematical notation could significantly enhance the applicability of academic writing automation, particularly in engineering and computer science domains where visual elements play crucial roles in research communication. Advanced evaluation frameworks that incorporate domain expert knowledge and subjective quality assessments represent another important research direction for developing more comprehensive and reliable assessment methodologies.

Practical applications of these findings include the development of domain-aware writing assistance tools that provide discipline-specific guidance and quality assessment. Educational applications could leverage these insights to create specialized training systems for academic writing across different fields. The integration of cross-disciplinary adaptation patterns into institutional writing support services could enhance research productivity while maintaining quality standards across diverse academic communities.

6. Acknowledgments

I would like to extend my sincere gratitude to Zhiyu Wang, Richard G. Baraniuk, and Andrew S. Lan for their groundbreaking research on scientific formula retrieval via tree embeddings as published in their article titled "Scientific formula retrieval via tree embeddings" in the 2021 IEEE International Conference on Big Data[21]. Their innovative approaches to academic content processing and embedding techniques have significantly influenced my understanding of advanced methodologies in scholarly text analysis and have provided valuable inspiration for my research in cross-disciplinary academic text generation.

I would like to express my heartfelt appreciation to Mengxue Zhang, Zhiyu Wang, Richard Baraniuk, and Andrew Lan for their innovative study on math operation embeddings for open-ended solution analysis and feedback, as published in their preprint "Math operation embeddings for open-ended solution analysis and feedback"[22]. Their comprehensive analysis of mathematical content representation and automated feedback generation has significantly enhanced my knowledge of academic text processing techniques and inspired my research in domain-specific language model adaptation for scholarly writing.

References:

[1]. Zhu, L., Yang, H., & Yan, Z. (2017, July). Extracting temporal information from online health communities. In Proceedings of the 2nd International Conference on Crowd Science and Engineering (pp. 50-55).

[2]. Zhu, L., Yang, H., & Yan, Z. (2017). Mining medical related temporal information from patients' self-description. International Journal of Crowd Science, 1(2), 110-120.

[3]. Zhang, D., & Jiang, X. (2024). Cognitive Collaboration: Understanding Human-AI Complementarity in Supply Chain Decision Processes. Spectrum of Research, 4(1).

[4]. Zhang, Z., & Zhu, L. (2024). Intelligent Detection and Defense Against Adversarial Content Evasion: A Multi-dimensional Feature Fusion Approach for Security Compliance. Spectrum of Research, 4(1).

[5]. Rao, G., Trinh, T. K., Chen, Y., Shu, M., & Zheng, S. (2024). Jump Prediction in Systemically Important Financial Institutions' CDS Prices. Spectrum of Research, 4(2).

[6]. Ju, C., & Trinh, T. K. (2023). A Machine Learning Approach to Supply Chain Vulnerability Early Warning System: Evidence from US Semiconductor Industry. Journal of Advanced Computing Systems, 3(11), 21-35.

[7]. Chen, Y., Ni, C., & Wang, H. (2024). AdaptiveGenBackend A Scalable Architecture for Low-Latency Generative AI Video Processing in Content Creation Platforms. Annals of Applied Sciences, 5(1).

[8]. Trinh, T. K., & Zhang, D. (2024). Algorithmic Fairness in Financial Decision-Making: Detection and Mitigation of Bias in Credit Scoring Applications. Journal of Advanced Computing Systems, 4(2), 36-49.

[9]. Raji, A. A. H., Alabdoon, A. H. F., & Almagtome, A. (2024, April). AI in Credit Scoring and Risk Assessment: Enhancing Lending Practices and Financial Inclusion. In 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS) (Vol. 1, pp. 1-7). IEEE.

[10]. Wu, J., Wang, H., Qian, K., & Feng, E. (2023). Optimizing Latency-Sensitive AI Applications Through Edge-Cloud Collaboration. Journal of Advanced Computing Systems, 3(3), 19-33.

[11]. Shih, J. Y., & Chin, Z. H. (2023, April). A Fairness Approach to Mitigating Racial Bias of Credit Scoring Models by Decision Tree and the Reweighing Fairness Algorithm. In 2023 IEEE 3rd International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB) (pp. 100-105). IEEE.

[12]. Zhu, C., Xin, J., & Zhang, D. (2024). A Deep Reinforcement Learning Approach to Dynamic E-commerce Pricing Under Supply Chain Disruption Risk. Annals of Applied Sciences, 5(1).

[13]. Zhu, C., Cheng, C., & Meng, S. (2024). DRL PricePro: A Deep Reinforcement Learning Framework for Personalized Dynamic Pricing in E-commerce Platforms with Supply Constraints. Spectrum of Research, 4(1).

[14]. Zhang, D., & Cheng, C. (2023). AI-enabled Product Authentication and Traceability in Global Supply Chains. Journal of Advanced Computing Systems, 3(6), 12-26.

[15]. Zhang, Z., & Wu, Z. (2023). Context-Aware Feature Selection for User Behavior Analytics in Zero-Trust Environments. Journal of Advanced Computing Systems, 3(5), 21-33.

[16]. Sun, M., Feng, Z., & Li, P. (2023). Real-Time AI-Driven Attribution Modeling for Dynamic Budget Allocation in US E-Commerce: A Small Appliance Sector Analysis. Journal of Advanced Computing Systems, 3(9), 39-53.

[17]. Zhang, S., Zhu, C., & Xin, J. (2024). CloudScale: A Lightweight AI Framework for Predictive Supply Chain Risk Management in Small and Medium Manufacturing Enterprises. Spectrum of Research, 4(2).

[18]. Zhang, S., Mo, T., & Zhang, Z. (2024). LightPersML: A Lightweight Machine Learning Pipeline Architecture for Real-Time Personalization in Resource-Constrained E-commerce Businesses. Journal of Advanced Computing Systems, 4(8), 44-56.

[19]. Kang, A., Xin, J., & Ma, X. (2024). Anomalous Cross-Border Capital Flow Patterns and Their Implications for National Economic Security: An Empirical Analysis. Journal of Advanced Computing Systems, 4(5), 42-54.

[20]. Zhao, Y., Zhang, P., Pu, Y., Lei, H., & Zheng, X. (2023). Unit operation combination and flow distribution scheme of water pump station system based on Genetic Algorithm. Applied Sciences, 13(21), 11869.

[21]. Wang, Z., Baraniuk, R. G., & Lan, A. S. (2021, December). Scientific formula retrieval via tree embeddings. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 1493-1503). IEEE.

[22]. Zhang, M., Wang, Z., Baraniuk, R., & Lan, A. (2021). Math operation embeddings for open-ended solution analysis and feedback. arXiv preprint arXiv:2104.12047.

[23]. Qi, D., Arfin, J., Zhang, M., Mathew, T., Pless, R., & Juba, B. (2018, March). Anomaly explanation using metadata. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1916-1924). IEEE.

[24]. Zhang, M., Mathew, T., & Juba, B. (2017, February). An improved algorithm for learning to perform exception-tolerant abduction. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

[25]. Yan, S. (2014). Design of Obstacle Avoidance System for the Blind based on Fuzzy Control. Netinfo Security.

[26]. Mo, K., Liu, W., Shen, F., Xu, X., Xu, L., Su, X., & Zhang, Y. (2024, May). Precision kinematic path optimization for high-dof robotic manipulators utilizing advanced natural language processing models. In 2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI) (pp. 649-654). IEEE.

[27]. Mo, K., Liu, W., Xu, X., Yu, C., Zou, Y., & Xia, F. (2024, May). Fine-tuning gemma-7b for enhanced sentiment analysis of financial news headlines. In 2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 130-135). IEEE.

[28]. Wu, S., Li, Y., Wang, M., Zhang, D., Zhou, Y., & Wu, Z. (2021, November). More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 2286-2300).

[29]. Wu, S., Wang, M., Li, Y., Zhang, D., & Wu, Z. (2022, February). Improving the applicability of knowledge-enhanced dialogue generation systems by using heterogeneous knowledge from multiple sources. In Proceedings of the fifteenth ACM international conference on WEB search and data mining (pp. 1149-1157).

[30]. Wu, S., Wang, M., Zhang, D., Zhou, Y., Li, Y., & Wu, Z. (2021, August). Knowledge-Aware Dialogue Generation via Hierarchical Infobox Accessing and Infobox-Dialogue Interaction Graph Network. In IJCAI (pp. 3964-3970).

[31]. Wang, M., Xue, P., Li, Y., & Wu, Z. (2021). Distilling the documents for relation extraction by topic segmentation. In Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16 (pp. 517-531). Springer International Publishing.

[32]. Eatherton, M. R., Schafer, B. W., Hajjar, J. F., Easterling, W. S., Avellaneda Ramirez, R. E., Wei, G., ... & Coleman, K. Considering ductility in the design of bare deck and concrete on metal deck diaphragms. In The 17th World Conference on Earthquake Engineering, Sendai, Japan.

[33]. Wei, G., Koutromanos, I., Murray, T. M., & Eatherton, M. R. (2019). Investigating partial tension field action in gable frame panel zones. Journal of Constructional Steel Research, 162, 105746.

[34]. Wei, G., Koutromanos, I., Murray, T. M., & Eatherton, M. R. (2018). Computational Study of Tension Field Action in Gable Frame Panel Zones.

[35]. Foroughi, H., Wei, G., Torabian, S., Eatherton, M. R., & Schafer, B. W. Seismic Demands on Steel Diaphragms for 3D Archetype Buildings with Concentric Braced Frames.

[36]. Zhu, L., Yang, H., & Yan, Z. (2017, July). Extracting temporal information from online health communities. In Proceedings of the 2nd International Conference on Crowd Science and Engineering (pp. 50-55).

[37]. Zhu, L., Yang, H., & Yan, Z. (2017). Mining medical related temporal information from patients' self-description. International Journal of Crowd Science, 1(2), 110-120.

[38]. Zhang, Z., & Zhu, L. (2024). Intelligent detection and defense against adversarial content evasion: A multi-dimensional feature fusion approach for security compliance. Spectrum of Research, 4(1).

[39]. Kuang, H., Zhu, L., Yin, H., Zhang, Z., Jing, B., & Kuang, J. The Impact of Individual Factors on Careless Responding Across Different Mental Disorder Screenings: A Cross-Sectional Study.

[40]. Cheng, C., Zhu, L., & Wang, X. (2024). Knowledge-Enhanced Attentive Recommendation: A Graph Neural Network Approach for Context-Aware User Preference Modeling. Annals of Applied Sciences, 5(1).

[41]. Wang, X., Chu, Z., & Zhu, L. (2024). Research on Data Augmentation Algorithms for Few-shot Image Classification Based on Generative Adversarial Networks. Academia Nexus Journal, 3(3).

[42]. Wang, M., & Zhu, L. (2024). Linguistic Analysis of Verb Tense Usage Patterns in Computer Science Paper Abstracts. Academia Nexus Journal, 3(3).

[43]. Guan, H., & Zhu, L. (2023). Dynamic Risk Assessment and Intelligent Decision Support System for Cross-border Payments Based on Deep

Reinforcement Learning. Journal of Advanced Computing Systems, 3(9), 80-92.

[44]. Zhu, L., & Zhang, C. (2023). User Behavior Feature Extraction and Optimization Methods for Mobile Advertisement Recommendation. Artificial Intelligence and Machine Learning Review, 4(3), 16-29.