Enhancing User Safety: Context-Aware Detection of Offensive Query-Ad Pairs in Multimodal Search Advertising

Gaurav Kumar*, Qiangjian Xi, Tanmaya Shekhar Dabral,
Hooshang Ghasemi, Abishek Krishnamoorthy, Danqing Fu,
Rui Min, Emilio Antunez, Zhongli Ding, Pradyumna Narayana
Google
***Corresponding author:** gauravkmr@google.com

Abstract

The proliferation of multi-modal online advertisements necessitates robust content moderation to ensure user safety, as offensive ad content can cause user distress and erode platform trust. This paper addresses the detection of content that becomes offensive only when a user's search query is paired with a specific ad, a context-dependent challenge that simple moderation often misses. Key challenges include the nuanced, multi-modal nature of ads, severe data scarcity and class imbalance due to the rarity of offensive content, and the high cost of human labeling. To overcome these limitations, we introduce a novel, context-aware detection framework centered on a large-scale, Multi-modal Teacher-Student Knowledge Distillation architecture. A powerful Gemini encoder-only "teacher" model distills its knowledge into a lightweight student model suitable for low-latency deployment. We enhance robustness using a novel graph mining technique to find rare offensive examples for training. For evaluation, we developed a highly accurate Automated Evaluation Model (AEM)—a separate, larger Gemini model utilizing Chain-of-Thought (CoT) reasoning—to rigorously assess performance in a live A/B test. Our results demonstrate that the proposed framework reduces the serving of offensive query-ad pairs by more than 80% compared to the baseline, while maintaining the efficiency required for real-time advertising systems that operate at a scale of over ≈ 100 billion query-ad pairs per day.

Disclaimer: This paper contains sentences and images that may be offensive. These examples are included solely for scientific analysis and do not reflect the views of the authors.

1 Introduction

The rapid expansion of e-commerce has transformed the global advertising landscape into a dynamic, multi-modal ecosystem (Kannan et al., 2017). While online advertisements bridge consumers and products (Verhoef et al., 2015; Williams, 2025), they introduce critical user safety challenges (Sadeghpour and Vlajic, 2021). Given the sheer scale where an ad's appropriateness is dictated by the user's query rather than intrinsic properties, manual moderation is impossible. Consequently, developing sophisticated automated systems is essential, as exposing users to offensive content erodes trust and undermines platform integrity (Gorwa et al., 2020).

A central difficulty in this domain is the **Challenge of At-Scale Multi-modal Contextual Safety**. Unlike traditional moderation, e-commerce advertising is inherently multi-modal (Yin et al., 2024), where offensiveness frequently emerges from semantic dissonance between a user's intent and the displayed ad (Kiela et al., 2021; Rathod, 2006). For instance, a benign ad for a "Night Out Dress" becomes highly inappropriate when served for the query "dresses for 8-year-old graduation." As this jarring experience emerges only at the millisecond-level intersection of a pre-approved creative and a live query, an extremely low-latency system capable of nuanced, context-aware reasoning is required.

The task is to determine if a given pair of (user query [text], ad creative [text, image]) is offensive. This presents distinct difficulties compared to general-purpose moderation:

- **Context Sensitivity and Multi-modality:** Offensiveness is dictated by the query-ad relationship. This dependency demands a multi-modal understanding of the user's intent, the ad's text, and its visual representation.
- **Data Scarcity and Class Imbalance:** Positive instances (offensive ads) are exceedingly rare (Saito and Rehmsmeier, 2015). This severe imbalance makes gathering reliable human ground-truth data difficult and

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 5: Industry Track, pages 475–482

costly, as raters struggle with focus and consistency amidst an overwhelming stream of non-offensive examples.

- **Viral Growth and Generalization:** Optimization systems may misinterpret clicks on shocking ads as relevance signals, potentially causing offensive content to go viral. A safety classifier must therefore generalize to filter novel offensive pairs proactively.

- **Low Latency Requirement:** As a user-facing system, inference must occur within strict millisecond constraints to ensure a seamless experience.

To address these challenges, we introduce a **Multi-modal Teacher-Student Knowledge Distillation Framework**. We first tackle data scarcity via a custom graph-mining strategy that expands a small seed set of offensive ads into a diverse dataset of "borderline" examples. We then fine-tune a state-of-the-art Gemini (Gemini Team, 2023) encoder, our "teacher", on this enriched data to achieve deep, contextual understanding. Subsequently, the powerful teacher's knowledge is distilled into a lightweight ResNet-based (Kaiming He and Sun, 2016) "student" model. The teacher generates a massive training set of pseudo-labels, allowing the student to learn from standard traffic distributions while mastering nuanced decision boundaries. Furthermore, we leverage a large Gemini-based model as an Automated Evaluation Model (AEM) for scalable online A/B testing (Yuan et al., 2024).

Our primary technical contributions are:

1. **Multi-modal Teacher-Student Knowledge Distillation:** A framework distilling a large Gemini teacher into a lightweight student model, balancing high accuracy with production-grade latency.

2. **Graph-Mining for Targeted Data Augmentation:** A custom pipeline that discovers "borderline" examples to robustly address severe class imbalance.

3. **Scalable Automated Evaluation:** Utilization of a fine-tuned Gemini model as a consistent automated rater for large-scale A/B experiments.

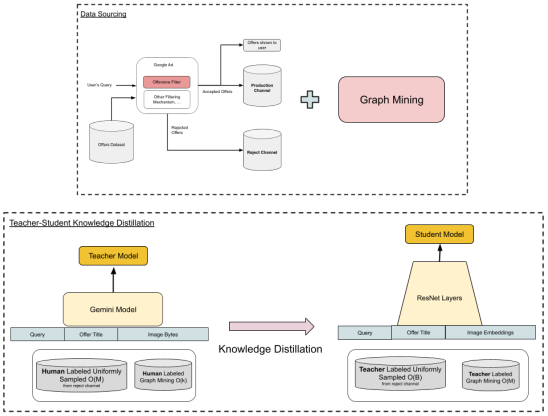

Figure 1: Teacher-Student Knowledge Distillation Framework. (Top) Graph mining pipeline augments data with borderline offensive examples. (Bottom) A teacher model distills knowledge into a lightweight student via pseudo-labels for low-latency deployment.

2 Data

Our data strategy addresses the challenges of scarcity and extreme class imbalance (He and Garcia, 2009) inherent in offensive content detection. This section details our approach to sourcing, annotation, preparation, and augmentation.

2.1 Data Sourcing

Training and testing data were sourced from two primary channels: *production* (ads displayed to users) and *reject* (ads filtered by existing systems). Sampling from both is essential for real-world robustness; relying solely on production data would blind the model to offensive examples currently being caught. We utilized both channels to ensure a comprehensive training distribution.

2.2 Human Annotation

To create a high-quality ground truth dataset, we utilized a rigorous human annotation process. Raters were presented with tuples of (query, ad image, ad title) and asked to assign an "offensiveness" label.

- **Rater Protocol:** Raters were provided with detailed guidelines and examples to distinguish between benign and offensive content, with a strong emphasis on the context provided by the user's query. The key task was to determine if displaying the specific ad in response to the given query was inappropriate or harmful.

- **Data Selection for Labeling:** Data selected for annotation included a mix from both production and reject channels, exposing raters to a wide spectrum of borderline and clearly offensive cases to serve as the seed for subsequent augmentation.

2.3 Data Preparation

Following sourcing and annotation, the data underwent a three-stage preparation phase to enhance label reliability:

1. **Denoising:** A majority-voting pipeline automatically harmonized conflicting labels from different human raters.

2. **Rater Scoring:** We periodically evaluated human raters on a "golden set" of unambiguous examples, selectively retaining labels only from consistent, high-scoring raters to mitigate label noise.

3. **Expert Review:** A targeted team reviewed likely mislabeled examples, such as high-confidence false negatives from preliminary models, to correct ambiguous cases.

Despite these protocols, positive instances (offensive ads) remained exceedingly rare. This extreme imbalance necessitated the data augmentation techniques described below.

2.4 Data Augmentation via Graph Mining

To counteract severe imbalance, we developed a custom graph mining augmentation pipeline to proactively identify likely offensive samples (Ma et al., 2023; Guo et al., 2022; Ren et al., 2024; Settles, 2009; Wang et al., 2025).

2.4.1 Pipeline to Generate Augmented Offensive Seed

This method expands an initial "offensive seed" of human-labeled ads. We construct a graph where nodes represent ads and weighted edges denote similarity, determined by both image embedding dot products and landing page similarity (Xiong et al., 2020). This approach effectively identifies offensive ads that are visually similar but use differing ad copy. A label propagation algorithm (Zhu and Ghahramani, 2002) spreads the "offensive" label to connected nodes (Figure 2). Ads exceeding a defined threshold form an augmented seed used to sample borderline examples.

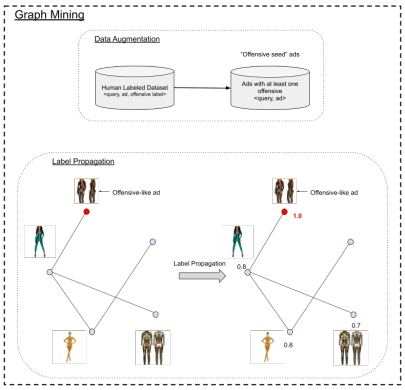

Figure 2: Graph Mining for Data Augmentation. An initial seed of human-labeled offensive ads is expanded via label propagation on an ad-similarity graph to generate a larger training dataset.

2.4.2 Graph Mining Dataset for Model Training

We utilized the augmented "offensive seeds" to identify borderline offensive query-ad pairs from the reject and production channel for training:

- **Teacher model:** Trained on several thousand human-annotated graph mining samples due to limited rating resources.

- **Student model:** Leveraging the scalability of our Teacher-Student Distillation approach, we trained Student model using O(millions) graph mining samples pseudo-labeled by the teacher.

3 Baseline Offensive Ads Model

Our initial baseline for detecting offensive query-ad pairs was a shallow, low-latency classifier designed for production efficiency. It operated on user query, ad title, and image embeddings (Jia et al., 2021), trained exclusively on human-annotated data without guidance of a large-scale teacher model. Text inputs were encoded via a bag-of-words model (Salton et al., 1975) using warm-started embedding tables. While efficient, this model's performance was fundamentally constrained by two data-centric challenges.

First, the subjective nature of "offensive" content caused significant label noise, stemming from inherent rater disagreements on borderline cases. Second, the rarity of genuine offensive pairings resulted in severe dataset imbalance, biasing the model towards the non-offensive majority. Standard mitigations like oversampling offered only

marginal benefits, often amplifying label noise present in the limited positive examples.

This potent combination of noisy labels and extreme data scarcity created an exceptionally difficult learning environment. The model was tasked with identifying a subtle pattern that was not only infrequent but also inconsistently defined. This data quality issue proved to be the primary bottleneck. We found that simply increasing model complexity could not compensate for this poor training signal; deeper architectures often became more confident in incorrect predictions derived from noisy, sparse data. Ultimately, performance remained capped by the quality of human annotations.

4 Multi-modal Teacher-Student Distillation Framework

Our proposed solution balances the trade-off between model accuracy and inference latency via a structured, multi-stage teacher-student framework. First, we enrich our dataset using a custom graph mining pipeline to identify rare, unsafe examples. Next, we fine-tune a large, multi-modal Gemini model to serve as a 'teacher' on this enriched data. We then utilize this teacher to generate pseudo-labels for billions of examples, distilling its knowledge into a massive, high-quality training set. Finally, we train an efficient 'student' model exclusively on these machine-generated labels for production deployment.

4.1 Teacher Model: Gemini Encoder-Only

This section details the Gemini encoder-only Teacher model, optimized for discriminative tasks, serving as the high-performance benchmark for student training.

4.1.1 Model Architecture and Initialization

We selected a Gemini Encoder-only architecture, engineered specifically for superior performance on scoring and classification tasks. We leverage transfer learning through a two-stage process:

1. **Foundational Initialization:** The encoder is adapted from a pre-trained, multi-modal Gemini decoder-only model. We modify the original causal attention mechanism to be bidirectional, enabling full context understanding. Final layer token embeddings are attention-pooled to generate a single, dense embedding suitable for classification (Vaswani et al., 2017; Sun and Lu, 2020).
2. **Task-Specific Fine-Tuning:** The model is fine-tuned on our human-labeled datasets for discriminative analysis. The final embedding is fed into a simple classification head (fully-connected layer with softmax) to output probabilities for predefined classes ("Offensive", "Not Offensive"). We utilized very large batch sizes to stabilize training given the extreme label imbalance.

4.1.2 Feature Engineering and Data Enhancements

Effective industrial models require robust feature representations to address two fundamental challenges: severe class imbalance and the semantic gap in interpreting multi-modal content.

4.1.2.1 Addressing Severe Data Imbalance

A canonical problem in safety systems is the extreme rarity of positive instances. Naively training on such skewed distributions biases models towards the majority class. We pursued strategic data sourcing, augmenting our primary training set with high-precision positive examples identified through custom graph-based mining (see Section 2.4). This uncovers semantically coherent clusters of offensive content missed by random sampling, improving minority class recall without compromising precision on legitimate content.

4.1.2.2 Multi-modal Feature Enhancement

Detecting offensive content requires interpreting subtle contextual interplay between text and visuals. We developed a two-pronged enhancement strategy:

Textual Feature Enrichment: To address query sparsity, we enriched primary textual inputs with auxiliary signals from the broader search context, such as search result titles. This grounds the model's understanding and reduces ambiguity.

Visual Feature Augmentation via Modality Translation: We introduced supplementary features that translate visual content into textual signals (Xu et al., 2015; Betker et al., 2023). A large Gemini decoder generates descriptive captions for ad images, which are evaluated by a dedicated safe-search system. The resulting classification scores are integrated as auxiliary input features. This provides the teacher model with explicit, pre-analyzed safety signals, helping it discern nuanced visual violations lacking explicit textual triggers.

4.2 Student Model

The student model employs knowledge distillation, trained solely on labels generated by the Gemini teacher (Li et al., 2024b). To close the "distillation gap" while adhering to strict production latency constraints, we widened the architecture rather than deepening it. We implemented a two-tower design where identical towers process input features (query, ad title, image embedding) in parallel. This allows simultaneous learning of complementary representations without the latency cost of serial layers. We further optimized tower depth by adding ResNet layers (Kaiming He and Sun, 2016) to maximize performance within our latency budget.

5 Automated Evaluation Model (AEM)

Scalable and consistent evaluation is a critical challenge in industrial safety (Li et al., 2024a; Pradel et al., 2024). Rating content for policy violations requires both high accuracy and interpretability crucial for transparency and debugging. Human rating often suffers from inconsistency and low throughput (Li et al., 2025; Doshi-Velez and Kim, 2017). To address this, we developed a multi-modal AEM using a large Gemini-based encoder-decoder model. Fine-tuned with Chain-of-Thought (CoT) prompting (Wei et al., 2023; Hsieh et al., 2023), it acts as both rater and reasoner, simultaneously producing classification labels and natural language justifications.

Validation against subject matter experts (SMEs) on a curated dataset of challenging, borderline examples demonstrated superior performance. In cases of disagreement adjudicated by a panel of policy experts, the AEM's original judgments aligned more frequently with the final consensus than the SMEs' labels. Consequently, the AEM served as our primary reliable rater for A/B testing, enabling high-fidelity evaluation at a scale prohibitive for human raters.

5.1 Rater Model Architecture

The AEM employs a hybrid architecture adapted from a standard Gemini decoder-only model. We created the encoder by modifying a copy of the decoder to use bidirectional attention. The encoder's final hidden state serves two parallel purposes: feeding a classification head for labeling, and providing context to the decoder for rationale generation.

Training leverages CoT data, where a larger Gemini model generates step-by-step "reasoning traces" for ground-truth labels. We utilize asymmetric backpropagation: generation loss updates both modules, while classification loss updates only the encoder. This integration acts as powerful regularization, forcing the encoder to learn representations rich enough for logical explanation and preventing reliance on superficial correlations.

5.2 Architectural Considerations for Rater vs. Teacher Models

A reader might question why a hybrid encoder-decoder architecture was chosen for the rater model, while an encoder-only architecture was used for the teacher model, especially since the teacher would also benefit from improved accuracy. We utilized a hybrid encoder-decoder for the rater, versus an encoder-only teacher, for two strategic reasons:

Resource Allocation: The teacher requires computational efficiency to label billions of examples. Conversely, the AEM evaluates a smaller, high-value set where accuracy and interpretability are paramount, justifying a larger, more complex architecture.

Bias Mitigation: Employing differing architectures for training (teacher) and evaluation (rater) minimizes shared inductive biases. A distinctly architectured, larger rater model provides a more objective assessment of performance, avoiding the blind spots inherent when rater and teacher share the same model structure.

6 Deployment

The Student model is deployed in a large-scale commercial advertising system, strategically positioned as one of the safety filters between the offer targeting stage and the final real-time auction.

6.1 Scale and Infrastructure

The deployed model, is served across a distributed infrastructure comprising ≈ 100 CPUs and ≈ 10 TPUs across ≈ 10 data center cells. This hybrid hardware approach allows us to balance throughput requirements with hardware availability.

In production, the model acts as a high-precision filter. Because it sits downstream of initial coarse safety filters and manual ad reviews, the incoming traffic is already largely clean. Consequently, the model maintains a highly selective filter rate of approximately 1 in 10,000 offers. This low rate highlights its role as a sophisticated final safeguard,

catching nuanced, context-specific violations that upstream systems miss.

6.2 Monitoring and Maintenance

To prevent silent performance degradation (e.g., due to concept drift in user queries or new ad trends), we implemented a robust monitoring pipeline:

1. **Drift Detection:** We continuously track the rolling mean of the model's prediction scores. Significant deviations trigger alerts for potential input distribution shifts.

2. **Throughput Monitoring:** We track Requests Per Second (RPS) to ensure the system meets strict Service Level Objectives (SLOs) during peak traffic requirements.

3. **Continuous Training (CT):** An automated CT pipeline periodically retrains the model on fresh data. We also employ a drift detection mechanism where the high-fidelity Teacher model periodically scores live traffic samples to audit the Student's ongoing performance.

7 Results

We validate our framework through both quantitative and qualitative evaluations.

7.1 Quantitative Results

Offline assessment on a held-out, human-labeled test set demonstrated significant improvement, with our student model achieving a >100% relative increase in Area Under the Precision-Recall Curve (AUCPR) over the baseline.

Subsequently, we conducted a large-scale live A/B experiment to measure real-world impact. Exposed to high volumes of unique query-ad pairs judged by our AEM, the student model reduced the serving of offensive pairs by >80% compared to the baseline while holding the false positive rate constant, confirming effective scaling of user safety.

7.2 Qualitative Analysis

We analyze the query **halloween costumes** to exemplify performance improvements. Figure 3 shows the baseline failing to filter an adult-themed latex mask. While acceptable for specific product searches, the ad is contextually inappropriate for this broad query—a mismatch the baseline missed. In contrast, our model correctly identified this nuance (Figure 4), filtering the potentially offensive result and replacing it with a benign alternative. Crucially, both models utilized identical input features, demonstrating our framework's capacity to capture complex contextual nuances without requiring additional signals.

Figure 3: Baseline model failing to filter a contextually offensive ad (highlighted red, blurred due to sensitive content) for the query "halloween costumes".

Figure 4: Proposed model successfully identifying and replacing the offensive ad with a benign alternative for the same query.

8 Conclusion

This paper introduced a novel Multi-modal Teacher-Student Framework to detect offensive query-ad pairs, addressing a critical challenge in search advertising safety. By distilling knowledge from a Gemini foundation model into a computationally efficient student, we achieved an 80% reduction in offensive ad serving relative to our baseline in production.

Our work offers three key contributions to industrial AI safety: utilizing graph mining to overcome extreme data imbalance; leveraging billions of teacher-generated pseudo-labels to bypass human annotation bottlenecks; and establishing an Automated Evaluation Model (AEM) for rigorous, scalable validation. Successfully deployed to process over $\approx$ 100 billion pairs daily, this framework provides a generalizable blueprint for engineering context-aware safety systems. It represents an adaptable paradigm that balances nuanced understanding with industrial scale, moving beyond

simple content filtering to set a new standard for responsible platform engineering.

9 Limitations

While effective, our system has limitations. First, it currently relies on upstream filters to catch the majority of gross violations; if those filters fail catastrophically, this model could be overwhelmed. Second, our current graph mining and teacher models are primarily optimized for English and major market languages; scaling nuance to all long-tail languages remains an ongoing challenge. Finally, the definition of "offensive" is culturally fluid and constantly evolving, necessitating continuous, expensive updates to our "golden" human-rated datasets to prevent model staleness.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. In *Computer Vision and Pattern Recognition*. Accessed via OpenAI paper release.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: A review of the landscape. *Internet Policy Review*, 9(4):1–20.

Hongwei Guo, Xin Wang, Yiding Wang, and Fuli Feng. 2022. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

H. He and E. A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Pfister, Yung-Sung Chung, S. Suda, R. Anil, and A. Bapna. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. 2016. "Deep Residual Learning for Image Recognition". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas.

Pallassana K Kannan and 1 others. 2017. Digital marketing: A framework, review and research agenda. *International journal of research in marketing*, 34(1):22–45.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Dawei Li and 1 others. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, Liwei Jiang, Nouha Dziri, Anne Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, and 1 others. 2025. SafetyAnalyst: Interpretable, Transparent, and Steerable Safety Moderation for AI Behavior. In *To be published*. Preprint available at arXiv.

Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. 2024b. Data generation using large language models for text classification: An empirical case study. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235. PMLR.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. 2023. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3649–3670.

F Philipp Pradel, Jon Roozenbeek, and Sander van der Linden. 2024. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120.

Ashish Rathod. 2006. A messaging system to handle semantic dissonance. Master's thesis, Rochester Institute of Technology.

Xubin Ren, Jiabin Tang, Dawei Yin, Nitesh Chawla, and Chao Huang. 2024. A survey of large language models for graphs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6616–6626.

Shadi Sadeghpour and Natalija Vlajic. 2021. Ads and fraud: A comprehensive survey of fraud in online advertising. *Journal of Cybersecurity and Privacy*, 1(4):804–832.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Peter C Verhoef, Pallassana K Kannan, and J Jeffrey Inman. 2015. From multi-channel retailing to omni-channel retailing: introduction to the special issue on multi-channel retailing. *Journal of retailing*, 91(2):174–181.

Zuhui Wang, Sandra Sajeev, Gaurav Mittal, Matthew Hall, and 1 others. 2025. Falcon: Fair active learning for content moderation. In *Computer Vision – ECCV 2024 Workshops*, pages 1–17. Springer. Preprint, publication details inferred.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jessica Williams. 2025. The omnichannel advantage: How online experiences strengthen the overall store. *Think with Google*. Accessed: 2025-07-31.

Jimmy Xiong, Matthijs Douze, Kaiming He, Kyunghyun Cho, Priya Goyal, and Hervé Jégou. 2020. Approximate nearest neighbor search on high dimensional data – experiments, analyses, and improvement. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 21768–21779.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, volume 37 of *PMLR*, pages 2048–2057.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Wei-Lin Yuan, Zhaode Wang, Hong-Ying Zan, Zhaohui Lin, Siyuan Bao, Yixin Wang, Jiacheng Liu, Yichi Zhang, Zhen Liu, Lisha Wang, and 1 others. 2024. A survey on leveraging large language models for natural language generation evaluation. *arXiv preprint arXiv:2401.07103*.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, Carnegie Mellon University, Pittsburgh.