



Article

From UGC to Brand Product Improvement: Mining Consumer Innovation Insights Across Social Media Platforms

Jiacheng Wang * and Qiang Wu

School of Management, University of Science and Technology of China, Hefei 230026, China;
qiangwu@ustc.edu.cn

* Correspondence: jiachengw@mail.ustc.edu.cn

Abstract

In the era of e-commerce, consumers' innovative insights are crucial for brands to improve their products and meet consumers' potential needs. Although user-generated content (UGC) platforms have accumulated a vast amount of consumer voices, brands often face the dilemma of "rich data but scarce insights": valuable consumer innovation insights are not only scarce but also expressed implicitly. Traditional methods have difficulty reliably identifying such insights in this task. Therefore, this study, for the first time, introduces a large language model (LLM)-augmented Siamese framework, aiming to improve the identification and generalization of consumer innovation insights in multi-platform UGC. Based on four Chinese social media platforms and a dataset of 133,538 comments, we conducted three experiments and performed a systematic comparison under five model configurations. The results show that our approach significantly outperforms traditional benchmarks and strong baselines in most experimental settings, while maintaining stable competitiveness in cross-platform tests. Finally, we showed how the identified innovation insights can be transformed into structured outputs for product planning, thereby facilitating the integration of consumer insights into the brand's innovation process.

Keywords: consumer innovation; social media insights; user-generated content; large language models; Siamese network

1. Introduction

In the age of e-commerce, the competitive advantage of brands increasingly relies on open, consumer-centered innovation ecosystems [1–3]. Consumers are shifting from passive buyers to active value co-creators, sharing their usage experiences, offering improvement suggestions, and expressing expectations for product innovation on social media user-generated content (UGC) platforms [4–6]. For brands, social media-based UGC platforms, because of their proximity to real usage scenarios and their real-time reflection of changing needs, have gradually become an important source for capturing innovative insights from both existing and potential consumers and enabling rapid product iteration [7,8].

Although the vast online comments on UGC platforms contain rich consumer innovation insights, accurately identifying these high-value insights in a high-noise environment is extremely difficult [9,10]. On the one hand, truly valuable innovation insights are statistically very scarce and are often buried under large amounts of irrelevant information and content. On the other hand, consumers' innovative ideas often show strong information stickiness [11], creating a significant "expression gap": unlike professional product managers, ordinary consumers lack sufficient domain knowledge and find it difficult to express



Academic Editor: Hua Pang

Received: 18 January 2026

Revised: 8 February 2026

Accepted: 9 February 2026

Published: 13 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

their thoughts as clear improvement suggestions [12,13]. Therefore, extracting innovative insights not only requires dealing with extreme class imbalance but also identifying implicit intentions in colloquial and diverse expressions [14], in order to transform scattered feedback into input that drives brands' product and service improvement.

In the field of UGC text mining, existing research has made significant progress, but there are still clear gaps in the task of "identifying high-value innovative insights." Most existing research focuses on sentiment analysis, topic modeling, defect identification, or conventional demand mining [15–17]. However, these goals often stop at understanding the current state of consumer feedback, without further identifying high-value insights that can drive iterative innovation in products and services. In fact, a specific and action-oriented innovation insight, given its strategic value, may far outweigh a vague consumer complaint. Although some studies have begun to move in this direction [18,19], they fail to account for the expression gap and information stickiness that may arise when consumers articulate their ideas, and thus do not address the challenges involved in identifying innovation insights. Moreover, these studies are often limited to a single platform or more structured scenarios, and their methods mainly rely on traditional machine learning or shallow deep learning models, making it difficult to effectively identify innovation insights in a multi-platform environment. Motivated by this gap, this paper focuses on a core research question: In a noisy, multi-platform UGC environment, how can we consistently and reliably identify scarce, implicitly expressed consumer innovation insights, while maintaining robustness and transferability under cross-platform differences in expression?

To address this research question, we propose a Siamese network augmented by large language model (LLM) [20], aiming to stably identify high-value consumer and prospective consumer innovation insights from multi-platform social UGC. To address the extreme sparsity of innovation insights and category imbalance, we fine-tune the pre-trained language model RoBERTa as the encoder [21], leveraging its general semantic priors to improve recognition with a small number of positive examples; at the same time, while retaining the core intent, we use generative AI to rewrite a small number of positive examples to increase the diversity of expressions in the minority class [22,23]. To address the "expression gap" caused by information stickiness, generative rewriting helps reduce ambiguity by turning fragmented and vague consumer expressions into clearer semantic variants; the Siamese network further strengthens the model's focus on the shared innovation insights behind different expressions through semantic consistency learning [24], thereby improving the stability of identifying implicit insights. To test the effectiveness of this framework under cross-platform differences in expression, we build a dataset from multiple social media platforms with clearly different styles and evaluate its robustness and transferability through multi-scenario experiments, including mixed training, platform-specific training, and leaving one platform out. As an application, this paper demonstrates how the identified innovative insights can be transformed into structured outputs for product planning, thereby supporting their use in the brand innovation process.

We conducted an empirical study in the context of the mobile service ecosystem. Based on a dataset of 133,538 manually annotated comments about HyperOS from four leading Chinese UGC platforms—Xiaohongshu, Douyin, Bilibili, and Weibo—we carried out a systematic empirical test. Focusing on the core task of identifying consumer innovation insights, we systematically compared five models under three settings: mixed data, platform-specific, and leaving one platform out. The overall results show consistent conclusions: models based on pre-trained representations significantly outperformed traditional machine learning and deep learning baselines across all settings [21,25]. Moreover, the LLM-augmented Siamese framework proposed in this paper further improved the recognition performance of scarce innovation insights in most scenarios and remained competitively stable in cross-platform

testing [20,22,24]. Overall, the above results provide empirical support for the validity and cross-platform robustness of the framework proposed in this paper.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the data and methodology. Section 4 reports the empirical results. Section 5 provides additional analysis. Section 6 discusses the academic implications and practical implications. Section 7 outlines limitations and directions for future research. Section 8 concludes.

2. Literature Review

This section reviews the relevant research on identifying high value innovation insights from UGC, mainly covering three aspects: the task landscape of consumer insight mining; key challenges such as sparsity, the expression gap, and cross-platform differences; and model and method cues related to this paper, including generative rewriting and Siamese networks.

2.1. Mining Consumer Insights

As an important carrier of consumer opinions in the era of e-commerce, UGC enables brands to capture consumer preferences and market information at an unprecedented scale [10,26,27]. As a result, “extracting consumer insights from UGC” has long been a core issue in the field of text mining [8]. Early studies mainly focused on sentiment analysis and opinion mining, using emotional and attitudinal information to explain or predict sales, satisfaction, and word-of-mouth diffusion, and combining methods such as topic modeling to identify the product attributes and usage scenarios that consumers care about, thereby addressing questions such as “how do consumers evaluate” and “what do they care about” [26,28,29].

Building on this, another important line of work is defect identification and diagnostic analysis, which aims to locate product malfunctions, service failures, and usage pain points in text to answer the question “Where is the problem?” [15,30]. Related methods have gradually evolved from traditional machine learning to deep learning and the fine-tuning of large language models, improving the accuracy of identifying sentiment, topics, and defect categories. This type of work is highly valuable for consumer experience monitoring and problem tracking, but its main focus still lies in explaining existing experiences and problems.

Furthermore, some studies have begun to focus on extracting requirements and suggestions, such as identifying suggestion sentences or product requirements [12,31,32], thereby moving feedback toward the “how to improve further” stage. However, in most research settings, the identified tasks mainly involve relatively clear and directly observable demand cues, rather than the rarer and more implicitly expressed high-value innovation insights [18,19]. Therefore, they usually do not need to deal with extremely sparse and highly imbalanced data scenarios. More importantly, different UGC platforms vary greatly in expression style and community context [33], which means that when the research target further shifts to “high-value innovation signal identification,” existing paradigms face new boundaries and challenges in both task definition and methodological capability.

2.2. Mining Innovation Insights

Identifying high-value innovative insights from large volumes of UGC is always challenging. The core reason is that such insights are often extremely sparse and expressed ambiguously [12]. According to information stickiness theory [11], innovation needs often depend heavily on context. Ordinary users tend to describe scenarios and pain points rather than directly offering actionable directions for improvement [12]. Therefore, innovation insights are often expressed in fragmented, indirect, and semantically ambiguous ways, which further increases the difficulty of stable identification and generalization [14].

At the methodological level, related research has expanded from relying on explicit cues to semantic modeling [14,25]. Early “suggestion mining” mainly depended on feature

engineering and rule-based patterns, which worked well for explicitly stated suggestions, but struggled to capture implicit needs [31]. With the development of deep learning, researchers began to use distributed representations to improve text representation. Models such as convolutional neural network (CNN) and long short-term memory (LSTM) have been applied to identify innovative content and have outperformed traditional bag-of-words methods [19]. However, in highly imbalanced and more freely expressed social media-based UGC, supervised classifiers are still prone to being dominated by the majority class, and their cross-platform recognition ability is also relatively limited [33].

In recent years, some studies have begun to explore alternative frameworks, such as anomaly detection, to identify sparse innovative ideas. For instance, innovative ideas are treated as outliers, and unsupervised autoencoders are used for identification [18]. These methods reduce reliance on labeled data, but they can be affected by the “rarity does not equal value” issue in social UGC. Outlier texts may simply be irrelevant information or even spam; moreover, equating innovation with anomalies also makes it difficult to capture underlying semantic consistency. In conclusion, relying solely on passive recognition is not sufficient for effective innovation mining. A better approach is to reduce ambiguity in language expression through LLM and combine it with the strong classification capability of pre-trained models [22,25].

2.3. LLM Augmentation and Siamese Network

Data augmentation techniques have become an important method for addressing data sparsity. In unstructured text classification, traditional augmentation methods such as synonym substitution can increase data volume, but they may alter semantics and even introduce noise [34,35]. Therefore, their ability to improve innovation classification under extreme class imbalance is limited, and they may even interfere with model learning.

In recent years, the development of LLMs has created more possibilities for data augmentation, thanks to their strong text generation capability. LLMs can generate high-quality text based on semantics rather than simple word substitution [22]. In the context of innovative insight mining, LLMs can help translate consumers’ vague expressions into clearer and more explicit statements to some extent. Building on this, we consider adopting an intent-preserving generative rewriting strategy to transform implicit expressions in the original comments into semantically clearer variants, in an attempt to address positive class scarcity and information stickiness.

However, if the model cannot distinguish between comments that use different expressions but convey the same core meaning, the effect of data augmentation will still be limited and may even lead to overfitting. Therefore, it is necessary to further introduce a Siamese network based on the Bidirectional Encoder Representations from Transformers (BERT) for metric learning [24]. For example, Reimers and Gurevych [20] proposed Sentence-BERT, which learns sentence semantics through a dual-tower structure and enables stronger robustness.

Although generative augmentation and Siamese learning have each been extensively studied [20,22,24], no research has yet integrated them into a unified solution for mining consumer innovation insights from UGC. The research framework proposed in this paper directly targets this gap.

3. Research Methodology

3.1. Data Collection

3.1.1. Research Context

This study selected Xiaomi HyperOS and its predecessor MIUI as the empirical subjects. The primary reason for this choice was to address a gap in the existing literature,

which mostly focuses on innovation in physical products but rarely examines complex software systems. With the rapid advancement of intelligent technologies and AI, product competitiveness is increasingly shaped by software-driven experiences. Therefore, exploring the OS innovation mechanism is of significant value.

In addition, this relevance is reinforced by Xiaomi's long-standing "user co-creation" culture and its collaborative innovation ecosystem [36,37], which provide rich, high-quality data for identifying consumer innovation insights. Moreover, HyperOS has gradually evolved into an ecosystem that covers the full "person-car-home" scenario and is deeply integrated with AI. This evolution not only reflects the future development trend of intelligent and connected systems, but also provides an ideal and challenging empirical setting for this study. In conclusion, HyperOS provides a typical and complex software scenario for validating the effectiveness and robustness of the identification framework proposed in this article.

3.1.2. Data Acquisition

In order to address the limitations of prior studies that relied on a single data source [18,19], and to test the transferability of the framework proposed in this paper under cross-platform differences in expression, we collected public online comments from multiple UGC platforms to better capture consumer innovation insights embedded in social media. As shown in Table 1, the four major UGC platforms we selected in China differ significantly in content form and community culture. This allowed us to construct a highly heterogeneous multi-platform corpus, reducing platform specific effects and enabling an evaluation of the model's cross-platform robustness. It also provides the necessary empirical basis for verifying the stability and broad applicability of this framework across different platform contexts. In total, we collected 210,965 public comments, spanning January 2012 to November 2025, with most comments posted in recent years.

Table 1. Comparative characteristics of data sources across four UGC platforms.

Platform	Interaction Focus	Content Modality
Xiaohongshu	Experience sharing	Image-text note
Weibo	Public discourse	Microblogging
Bilibili	Interest interaction	Long-form video
Douyin	Instant expression	Short video

3.1.3. Data Preprocessing

To support the subsequent analysis, we applied standardized preprocessing to the data [38]. First, we removed non-semantic elements such as emoticons and web links from the text. Second, given that valuable innovative comments are usually not overly brief [18], we filtered out comments with fewer than 10 characters and deduplicated the data. Ultimately, we obtained 133,538 valid comments, and the relevant information is shown in Table 2. The table also shows substantial differences in average comment length across platforms: comments on Bilibili are the longest (31.70), while those on Douyin are the shortest (22.53), which to some extent reflects differences in expression styles across platforms.

Table 2. Descriptive statistics of data distribution across platforms.

Platform	Raw Count	Valid Count	Avg. Length
Xiaohongshu	55,999	31,290	24.01
Weibo	31,767	27,501	23.42
Bilibili	52,228	36,270	31.70
Douyin	70,971	38,477	22.53
Total	210,965	133,538	25.55

3.2. Constructing the Ground Truth

To validate the proposed model, we first established reliable benchmark labels. Drawing on relevant studies [12], we designed and implemented a rigorous manual annotation process.

3.2.1. Concept Definition and Annotation Process

We define comments with innovative insights ($y = 1$) as those that contain specific, actionable ideas for improvement. A comment is labeled as innovative if it meets any of the following conditions: clearly indicates new functional requirements that are currently absent; proposes forward-looking suggestions for existing functions; or indirectly points to a direction for improvement by describing pain points in specific usage scenarios. Referring to the defect classification framework of Abrahams et al. [15], we clearly treat general problem reports and emotional expressions as not belonging to innovative insights.

We invited three master's students majoring in innovation management to perform the annotations. To ensure annotation quality, the process was divided into three stages: first, we conducted standardized training and a pilot annotation; second, the three students independently annotated the entire dataset; finally, we conducted a consistency check. From the independently annotated data, we randomly selected 10% of the samples and computed Fleiss' kappa [39], which was 0.87, indicating very high inter-rater agreement. We then used majority voting to determine the final labels.

3.2.2. Sparsity and Heterogeneity in Data Distribution

The annotation results show that highly valuable innovative insights are extremely scarce. Among the 133,538 valid comments, only 680 were identified as innovative, accounting for 0.51% of the total. This extreme class imbalance further highlights the "data abundant but insights scarce" problem.

As shown in Figure 1, the distribution of innovation insights differs markedly across platforms. The innovation rate on Weibo is the highest, reaching 1.35% ($n = 372$), which may be because consumers tend to leave comments and provide feedback under official enterprise accounts or executives' Weibo posts. In contrast, the innovation rates on Douyin (0.24%, $n = 91$) and Bilibili (0.28%, $n = 101$) are much lower. This further highlights heterogeneity and differences in expression styles across platforms. When brands seek innovative consumer insights to support improvements in their products and services, they must pay attention to these cross-platform differences.

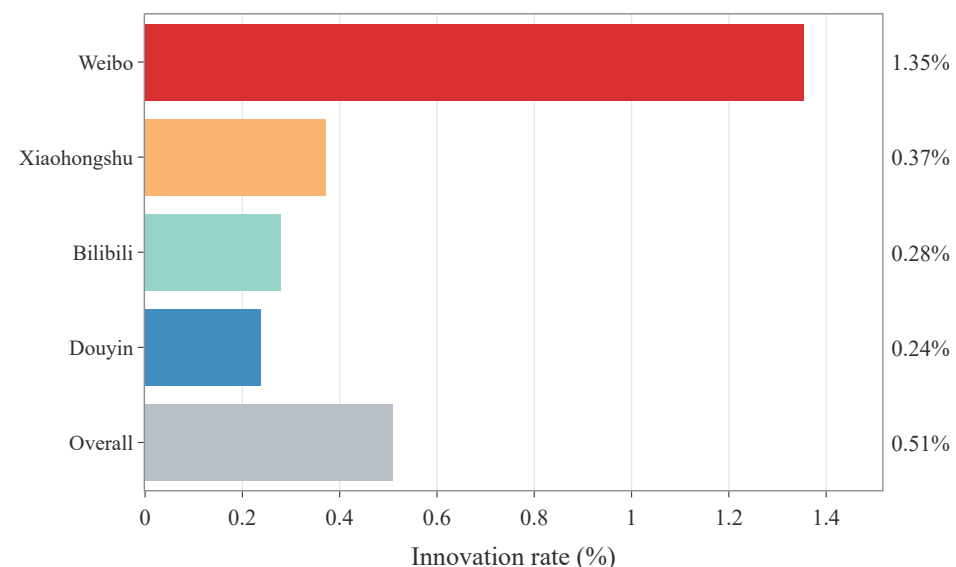


Figure 1. Distribution of innovation rates across four UGC platforms.

We further plotted the kernel density estimation (KDE) curves, as shown in Figure 2. The figure shows that innovative and non-innovative comments follow markedly different distributions. Innovative comments (red curve) exhibit a clear long-tail pattern, with an average length of 60.2 characters, whereas non-innovative comments (blue curve) are much more concentrated at shorter lengths, with an average of 25.4 characters. Compared with simple feedback and emotional expressions, innovative comments often require richer context and a more forward-looking perspective, and are therefore harder to express in a few words. This pattern is consistent with the theoretical view of “information stickiness” [11].

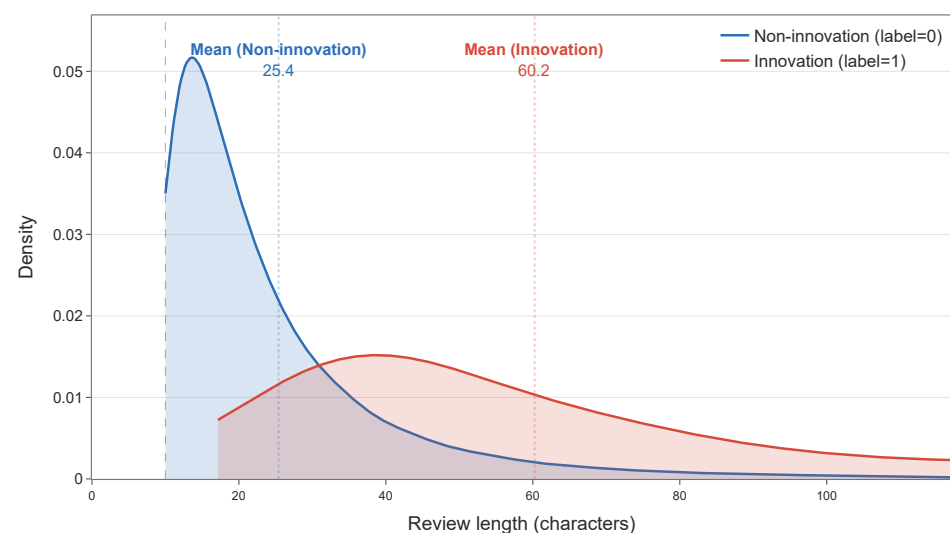


Figure 2. Kernel Density Estimation of review length: Innovation vs. Non-innovation.

3.3. Model Development

To systematically address the dual challenges of sparsity and information stickiness identified in Section 3.2, we developed five model variants and compared their performance.

3.3.1. Methodological Spectrum: Baseline Models

We used three representative benchmark models (Models 1–3) for comparison. Model 1 is logistic regression, a classic machine learning method; Model 2 is LSTM, a classic deep learning method; and Model 3 is based on the Transformer architecture and represents a pre-trained language model.

Model 1 applies logistic regression to term frequency–inverse document frequency (TF–IDF) features [25,40], constructing a high-dimensional vector space based on the frequency of explicit keywords. Although it is computationally efficient, it is constrained by the “bag-of-words” assumption and cannot capture semantic context or understand complex consumer intentions.

Model 2 uses a LSTM network initialized with static Word2Vec embeddings [41,42]. Although it can capture sequential information, its word representations are context-independent, making it difficult to adapt when meanings shift with context.

Finally, Model 3, as a representative pre-trained language model, is implemented based on the standard RoBERTa architecture [21]. We built a strong benchmark model by fine-tuning the RoBERTa-wwm-ext backbone with a conventional single-tower classification head [43]. It should be noted that, for the specific task of innovative insight recognition, studies that directly introduce pre-trained models are still scarce. Therefore, Model 3 not only provides a reliable performance reference for this task but also lays an important model foundation for the fusion framework proposed in this paper.

3.3.2. The Proposed Framework: LLM-Augmented Siamese Framework

To address the limitations of the aforementioned baseline methods, we introduce Model 5, the LLM-augmented Siamese framework. As shown in Figure 3, this framework consists of two stages. First, a large language model rewrites innovative comments in an intent-preserving manner to expand the range of expressions in the minority class. Second, a shared-parameter Siamese RoBERTa encodes the original text and the rewritten text in parallel, and similarity constraints encourage the model to focus more on core semantics rather than surface wording, thereby improving the stability of identifying implicit innovative insights.

Additionally, to rigorously isolate the contribution of the generative augmentation module, we construct an ablation variant denoted as Model 4. This model retains the Siamese learning architecture (Phase 2) but excludes the generative augmentation process (Phase 1), training solely on the original comments.

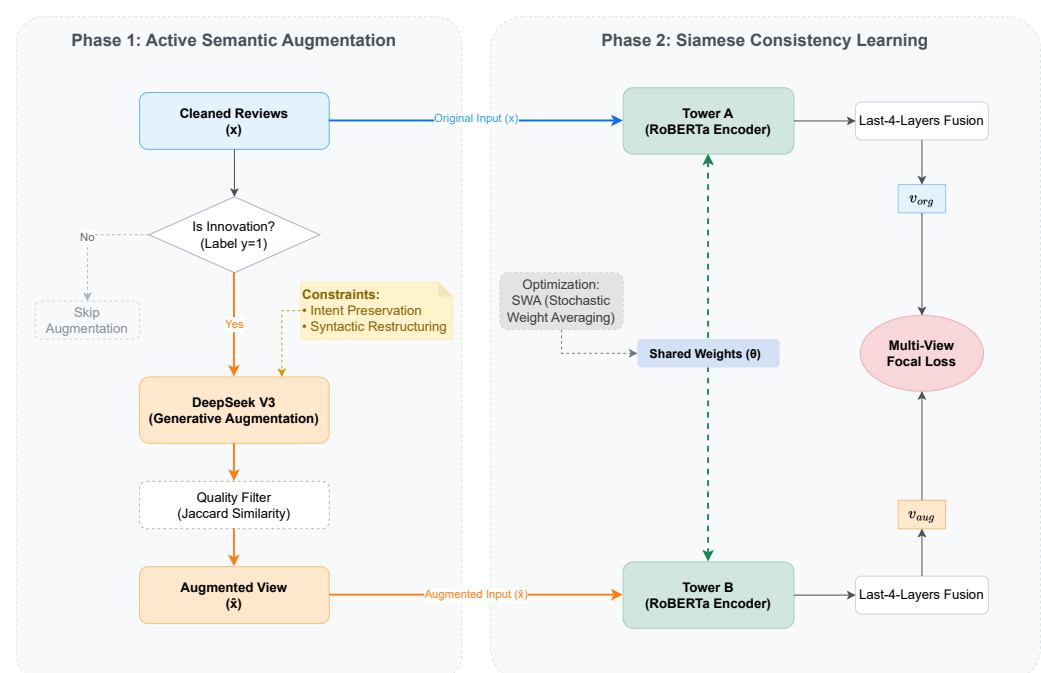


Figure 3. Schematic overview of the LLM-Augmented Siamese Framework.

3.3.3. Phase I: Active Semantic Augmentation via Cognitive Proxies

Innovation mining is fundamentally constrained by the expression gap. Conventional augmentation methods such as synonym replacement often ignore context and can cause semantic drift [34,35]. To narrow this gap, we treat LLM as a cognitive proxy [44] that helps rephrase consumer intent without changing what is meant.

Using DeepSeek V3 [45], we define a generation function $\mathcal{G}(x | c)$ that rewrites a comment x under pragmatic constraints c . The goal is to produce intent-preserving variants that make implicit insights easier to learn. This rewriting is applied only to comments labeled as innovation insights ($y = 1$). We apply LLM rewriting only to the training set.

To reduce unfaithful generations, we filter candidates with Jaccard similarity [46]. A generated comment $\hat{x}$ is kept only if $\tau_{min} < \text{Jaccard}(x, \hat{x}) < \tau_{max}$. The lower bound τ_{min} helps prevent drift from the original intent, while the upper bound τ_{max} avoids near-duplicates and encourages expression diversity. As shown in Table 3, the retained variants make latent intent more explicit and closer to structured feature requests.

Table 3. Sample Innovation Insights and Augmented Semantic Variants.

Original Review (x)	Augmented Variant 1 ($\hat{x}_1$)	Augmented Variant 2 ($\hat{x}_2$)
As a hearing impaired user, [...] the experience is terrible. The haptics setting is not vibration for us. [...] The normal function should support message classification. [...] I hope the developer sees this.	Suggestion to overhaul the haptic feedback system by implementing specific vibration patterns for distinct notification categories (IM apps) and introducing granular intensity modulation to ensure accessibility.	An ideal solution would be an accessibility option that lets users adjust vibration duration, frequency, and strength levels, so tactile alerts are easier to feel for people who are hearing-impaired.
Xiao Ai can't distinguish the owner's voice from others, when I open it while watching TV, she proactively identifies the TV sound and talks nonsense [...] I didn't see any option in the settings.	The current voice activation often picks up background noise and misses my commands. It would be really helpful to add "Owner Voiceprint Isolation" to filter out sources like TV audio.	Xiao Ai doesn't handle background noise well. I suggest adding a strict "Owner-Only" mode that verifies the owner's voice, so it won't be triggered by TV, videos, or other media nearby.
When can you release an automation scheme where if the watch judges that a person is asleep and the phone screen is off, it automatically turns off the lights?	Inquiry about a more complex automation setup: triggering an automatic lights off routine when my wearable detects I'm asleep and my phone screen is locked.	I'd like a smoother ecosystem feature where the system controls the lights by combining my watch's sleep status with my phone's screen lock status.

Note: x denotes the raw user input; $\hat{x}_1$ and $\hat{x}_2$ denote the semantic variants generated by DeepSeek V3.

3.3.4. Phase II: Semantic Consistency Learning via Siamese Network

To distinguish high-value innovative ideas from general noise, we employ a Siamese network [20]. Unlike standard classifiers, this dual-tower structure learns a metric space where inputs with similar semantic intent are clustered together, regardless of their surface-level linguistic variations.

Hierarchical Feature Fusion: We utilize RoBERTa-wwm-ext as the backbone encoder [43]. Recognizing that innovation signals often rely on a combination of specific syntactic cues and deep semantic content, we propose a Last-4-Layers Fusion strategy. Let $H^{(l)}$ denote the hidden states of the l -th transformer layer. We compute the final sentence representation $v \in \mathbb{R}^d$ by averaging across the last four layers:

$$v = \frac{1}{4} \sum_{l=L-3}^L \left(\frac{1}{N} \sum_{i=1}^N h_{i,l} \right) \quad (1)$$

This hierarchical fusion ensures that the model captures both the abstract concept of innovation and subtle pragmatic nuances [47].

Weight Sharing Mechanism: The architecture consists of two identical sub-networks sharing the same parameters θ . During training, the original review x and its augmented variant $\hat{x}$ are processed in parallel:

$$v_{org} = f_{\theta}(x), \quad v_{aug} = f_{\theta}(\hat{x}) \quad (2)$$

This mechanism acts as a regularizer, forcing the encoder to map lexically diverse but semantically equivalent texts into a consistent decision space, effectively learning the invariant core of the innovation idea.

3.3.5. Optimization Strategy

To transform the semantic representations into innovation probabilities, we introduce a shared classification head $g_{\phi}(\cdot)$. The final prediction is given by $p(\cdot) = g_{\phi}(f_{\theta}(\cdot))$.

Multi-View Auxiliary Supervision: Given the extreme class imbalance, we employ the Focal Loss [48] as the core objective function. Let p_t be the model's estimated probability for the ground-truth class. The Focal Loss is defined as $\mathcal{L}_{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t)$. We further introduce a multi-view auxiliary supervision objective. The total loss $\mathcal{L}_{Total}$

integrates supervision from both the primary view (original data x) and the auxiliary view (augmented data $\hat{x}$):

$$\mathcal{L}_{Total} = \mathcal{L}_{FL}(p_t(x)) + \lambda \cdot \mathcal{L}_{FL}(p_t(\hat{x})) \quad (3)$$

where λ is a hyperparameter balancing the contribution of the augmented data. Finally, Stochastic Weight Averaging (SWA) is applied to flatten the loss landscape and enhance generalization [49].

3.4. From Insight Detection to Strategic Analysis

In the previous section, we developed multiple model schemes and conducted comparative evaluations to address the challenge of identifying consumer innovation insights. However, to connect the algorithm outputs with actionable product iteration strategies, we conducted post-analysis, including semantic clustering and value prioritization mechanisms.

3.4.1. Semantic Clustering and Taxonomy Generation

To uncover the potential thematic structure of innovative comments, we used the semantic representations learned by Model 5 and clustered these vectors. We then selected representative samples from each cluster by choosing the comments closest to the cluster centroid. We invited three industry experts, including an internet product manager and two business analysts, to summarize these comments into a hierarchical thematic structure with primary and secondary categories, in order to interpret the semantic meaning of the clusters.

3.4.2. Value-Urgency Prioritization Mechanism

We also introduced a two-dimensional scoring mechanism, converting the classified innovative insights into a clear priority roadmap. We continued to invite three experts to independently rate each innovative comment from 1 to 10. The assessment considered two dimensions: strategic value and development urgency [50–52]. Strategic value reflects the potential impact of an innovative idea on the product's long-term differentiation and competitive advantage, while development urgency reflects the urgency of consumer demands and the potential losses from missed market opportunities [53].

4. Empirical Analysis and Results

4.1. Evaluation Metrics and Experimental Design

4.1.1. Evaluation Metrics

Because the proportion of positive cases in this study is extremely low, accounting for only about 0.51%, conventional metrics such as accuracy can be misleading. Even if all samples are predicted as the majority class, accuracy can still exceed 99.4%, but this has no practical significance [54]. Therefore, we selected precision, recall, and F1 score as the core evaluation metrics. From a practical perspective, precision ensures that resources are not wasted on low-value content, while recall ensures that scarce, high-value innovative insights are captured as fully as possible.

To further evaluate overall model performance, we also introduced the area under the precision–recall curve (PR-AUC). Compared with the area under the receiver operating characteristic curve (ROC-AUC), PR-AUC is more informative under extreme class imbalance because it focuses more on the model's performance on the positive class, making it more suitable for assessing performance in severely imbalanced tasks [55].

4.1.2. Experimental Design Scenarios

To comprehensively evaluate the robustness and generalization capabilities of the proposed framework, we designed three experimental scenarios. Scenario I uses merged

data from four platforms for unified training and testing. It assesses the model's overall performance in identifying innovative insights from mixed, heterogeneous comments. This scenario serves as the core benchmark for evaluating the proposed method relative to baseline models. Scenario II trains and tests separately on each platform. It examines the model's adaptability to distinct community cultures and expressive characteristics across UGC platforms. Scenario III adopts a rigorous leave-one-platform-out (LOPO) validation protocol [56]. Training uses data from any three platforms, and testing is conducted on the remaining platform. This setup evaluates cross-platform generalization when the model faces substantial domain differences.

4.1.3. Model Configuration and Implementation

We compare five model groups across three experimental settings (mixed training, single-platform training, and leave-one-platform-out testing). Model 1 uses term-frequency-inverse document frequency (TF-IDF) features with an ℓ_2 -regularized logistic regression classifier; we tune the regularization strength via grid search and select the operating threshold by scanning the validation set. Model 2 is a Word2Vec initialized LSTM classifier (two layers), trained under a matched budget and likewise paired with validation-based threshold selection to improve performance under severe class imbalance. Models 3–5 are implemented in HuggingFace with RoBERTa-wwm-ext and trained for a unified budget of 12 epochs using focal loss ($\gamma = 2.0$) and AdamW; we further apply stochastic weight averaging (SWA) to improve training stability. To ensure strict comparability, all models reuse the same data split (70%/15%/15% for train/validation/test) across models and experimental settings.

4.2. Analysis on the Aggregated Dataset

Table 4 summarizes the performance metrics of the five models in Scenario I. The empirical results clearly demonstrate the proposed framework's capability in mining sparse innovative insights.

Table 4. Performance comparison of different models on the test set.

Model Configuration	Precision	Recall	F1 Score	PR-AUC
Baselines				
Model 1	0.5347	0.5294	0.5320	0.5607
Model 2	0.6762	0.6961	0.6860	0.6909
Model 3	0.8602	0.7843	0.8205	0.8958
Proposed Framework				
Model 4	0.9000	0.7941	0.8437	0.8980
Model 5	0.8491	0.8824	0.8654	0.9291

Based on the performance of each model, the traditional benchmark models show clear limitations. Model 1 is constrained by shallow text representations and struggles to capture semantic context effectively, with an F1 score of only 0.5320. Model 2 improves by introducing sequence modeling capability, with its F1 score rising to 0.6860. However, its overall discriminative ability remains limited, with a PR-AUC of only 0.6909. These results indicate that methods that mainly rely on static word vectors or local features cannot cope with the inherent ambiguity and diversity in consumer expressions.

In contrast, Model 3 achieves stronger recognition capability based on a pre-trained language model, providing a much more robust baseline for innovative insight identification. It reaches an F1 score of 0.8205, indicating that contextual representations have an advantage in understanding implicit innovation insights. Model 4 introduces a Siamese network on top of Model 3, enabling more effective separation of confusing samples and

improving precision: Model 4 achieves a precision of 0.9000, compared with 0.8602 for Model 3.

Model 5 achieves the best overall performance, with an F1 score of 0.8654 and a PR-AUC of 0.9291. In addition, Model 5 raises recall to 0.8824. Compared with Model 3, it identifies approximately 9.8% more high-value insights, and compared with Model 4, it identifies approximately 8.8% more. This result indicates that, in a multi-platform data environment with substantial style differences, Model 5 maintains stronger recognition performance than traditional methods and strong benchmark models, thereby more reliably capturing high-value innovation insights that might otherwise be overlooked amid extensive noise.

This paper further presents the PR curves of each model, as shown in Figure 4, to visually compare performance differences among the five models. The curve of Model 5 consistently dominates the other benchmark models in the high-recall region. This pattern also illustrates the trade-off between R&D efficiency (measured by precision) and opportunity coverage (measured by recall). By adjusting the classification threshold along this curve, brands can balance efficiency and comprehensiveness in data mining, thereby tailoring innovation insight mining strategies to different development stages.

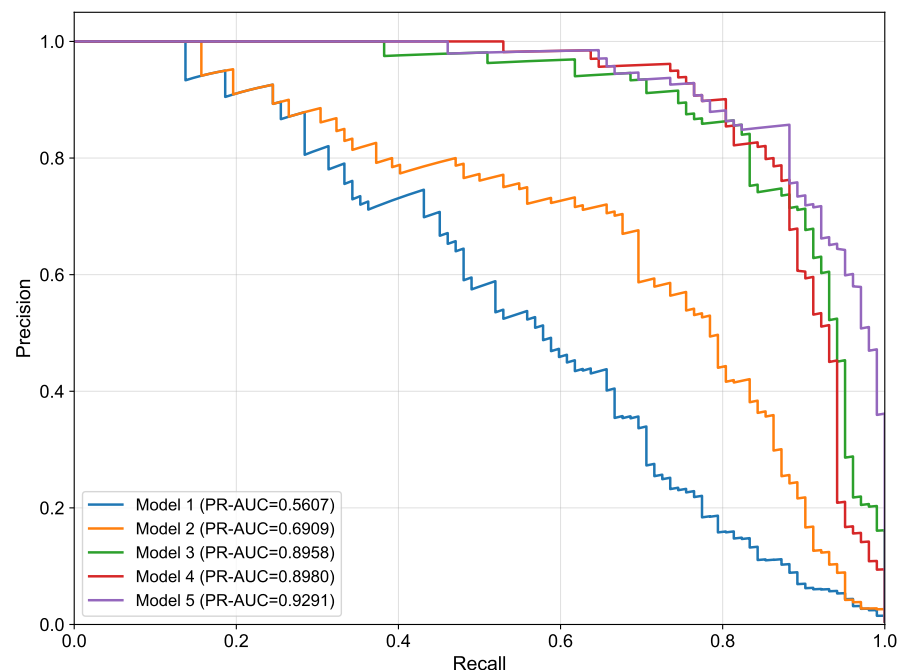


Figure 4. Precision-Recall curves for all models on the test set.

4.3. Single Platform Analysis

To evaluate the applicability of the proposed framework across UGC platforms, we conducted separate assessments for each platform, corresponding to Scenario II. Figure 5 reports the F1 scores and PR-AUC values for all five models across the four data sources. Detailed metrics, including precision and recall, are provided in Appendix A Table A1.

In the single-platform test of Scenario II, the performance of all models was lower than in the mixed-training scenario. This was mainly due to the relatively small size of each platform-specific dataset. However, Model 5 proposed in this study still demonstrates more stable recognition performance on most platforms.

On fragmented and life-oriented platforms such as Xiaohongshu and Douyin, Model 5 maintains a leading position and is more stable overall. Taking Xiaohongshu as an example, Model 5 achieves an F1 score of 0.679 (Model 3: 0.667; Model 4: 0.600) and a PR-AUC of 0.749 (Model 3: 0.719; Model 4: 0.652); on Douyin, Model 5 still ranks first with a PR-AUC

of 0.812. These results indicate that the generative enhancement mechanism is more helpful for identifying implicit innovation insights, whereas the gain from simply introducing the Siamese network is relatively limited.

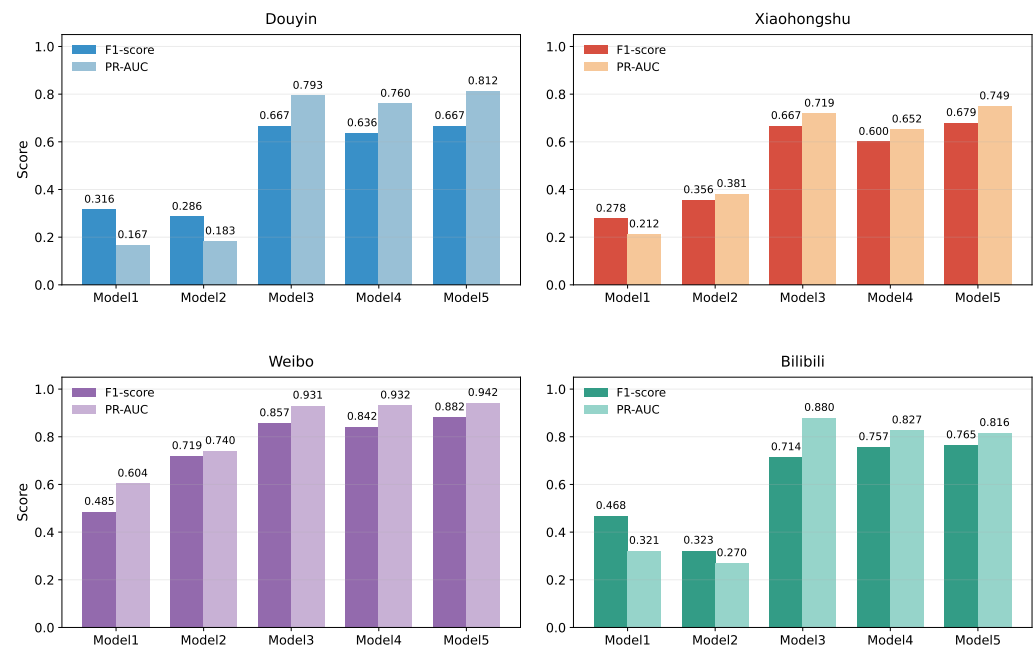


Figure 5. Performance comparison across four individual platforms.

On platforms such as Weibo, which focuses on public discussions, and Bilibili, which centers on specific interests, the models show clearer performance trade-offs. On Weibo, Model 4 achieves a PR-AUC of 0.932, indicating better ranking stability; on Bilibili, Model 3 attains the highest PR-AUC (0.880), while Model 5 delivers the best F1 score (0.765). This suggests that across different community contexts, the Siamese network helps strengthen judgments of semantic consistency, whereas the full LLM-augmented Siamese framework offers a greater overall advantage in balancing recall and accuracy.

It is worth noting that all models perform best on Weibo, mainly because positive samples are more abundant. In contrast, across the other three platforms, a clear pattern emerges: the sparser the data, the more pronounced the disadvantage of traditional methods, and the more prominent the advantage of pre-trained language models.

In conclusion, the single-platform results indicate that Model 5 is more robust under limited data and maintains strong performance across platforms with different content styles and varying degrees of class imbalance.

4.4. Leave One Platform out Analysis

To rigorously evaluate the generalization ability and external validity of the proposed framework, we further conducted a LOPO experiment [56]. Under this protocol, the model was trained on data from three platforms and then evaluated on a completely unseen fourth platform. Figure 6 summarizes the performance of the five models. Full metrics, including precision and recall, are reported in Appendix A Table A2.

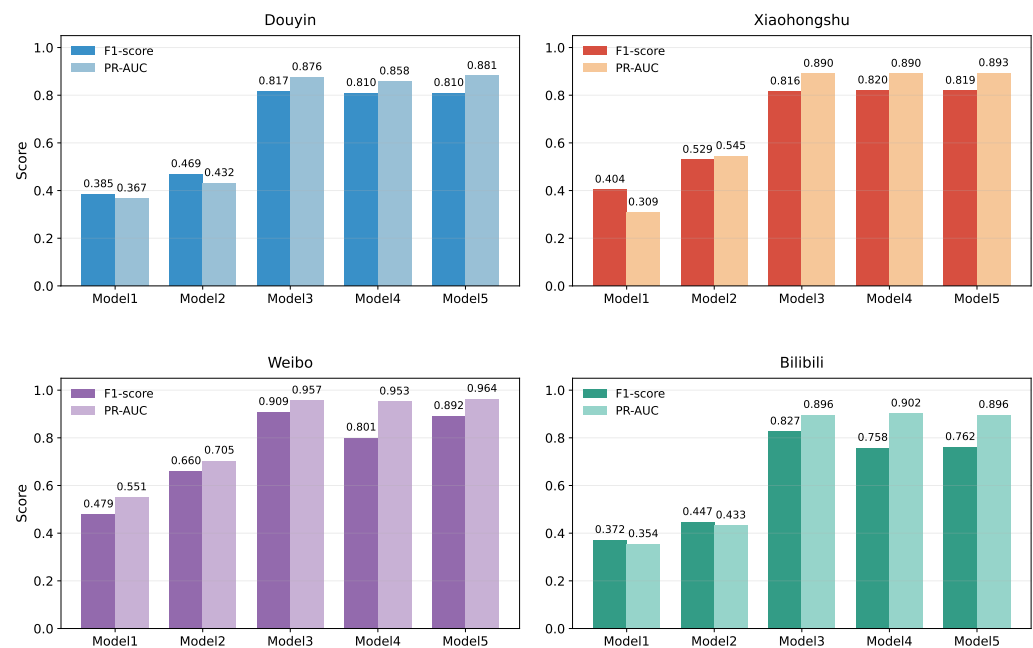


Figure 6. Cross-platform generalization performance under the LOPO protocol.

The results show that, when facing domain differences, there are substantial performance gaps among these models. The traditional benchmark models (Model 1 and Model 2) experience a notable performance decline compared with the other models under the same protocol. For example, on Xiaohongshu, Model 1 achieves an F1 score of only 0.404, while Model 2 reaches 0.529. This performance collapse indicates that shallow statistical features and static embeddings cannot capture invariant semantic structures of innovation insights, making them ineffective for cross domain transfer. Turning to the other three models, Model 4 relies only on the Siamese network and does not incorporate any enhancement, and its performance is usually weaker than Model 3 and Model 5. This suggests that relying solely on the Siamese network may lead to overfitting to source-specific patterns, thereby limiting generalization to unseen domains.

The comparison between Model 3 and Model 5 reveals the boundary conditions of generative enhancement. On Xiaohongshu, Model 5 achieves an F1 score of 0.819, slightly higher than Model 3's 0.816; on Douyin, the two models perform similarly, whereas on Weibo and Bilibili, Model 3 shows slightly better generalization. For example, on Bilibili, Model 3 attains an F1 score of 0.827, higher than Model 5's 0.762. This may be because LLM-generated samples still carry subtle semantic biases from the source training platforms, while Model 3's representations, grounded in general pre-training knowledge, are more stable.

These findings provide clear guidance for management practice: when the target platform has historical data for fine-tuning, the framework proposed in this study can achieve the best results; when exploring new channels without labeled data, the standard pre-trained model remains a reliable tool for preliminary exploration.

5. Additional Analysis

After conducting a systematic evaluation of the performance of five models across three experimental scenarios, this paper further demonstrates how the identified innovative insights can be presented in a more manageable and usable format. Specifically, we summarize the discovered insights into structured outputs that can support product planning.

5.1. Uncovering Latent Innovation Themes

Based on the 680 innovative comments that were manually labeled previously, we first used the trained Model 5 encoder to extract high-dimensional semantic vectors. We then applied the K-means algorithm to cluster these semantic representations and reveal the underlying thematic structure [57]. Using t-SNE [58], we performed dimensionality reduction and visualized the clustering results in a two-dimensional space.

Figure 7 presents the semantic clustering results and their visualization. The figure shows that innovative insights form several distinct clusters in semantic space. Based on this, we selected representative core comments near each cluster center and used DeepSeek V3 to make preliminary topic judgments. We then invited three experts from the internet industry to review and refine the results, and finally determined the topic labels for each cluster.

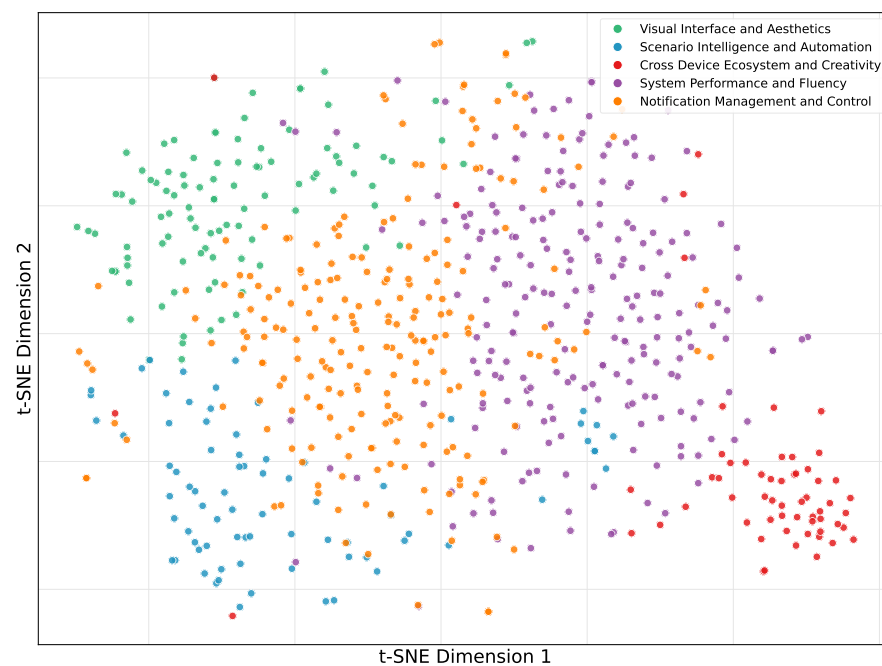


Figure 7. t-SNE visualization of the semantic space for innovation insights.

Figure 7 shows that “Cross Device Ecosystem and Creativity” forms a distinct cluster, which is directly related to multi-device collaboration and ecosystem linkage. Because these topics usually have a higher threshold for expression and often point to the frontier of product development, this cluster is relatively small. In addition, “System Performance and Fluency” and “Notification Management and Control” reflect basic consumer experiences that can be directly perceived. They account for the largest share of the sample and form the largest cluster. Together, these five categories form a demand-structure framework that can help managers map scattered innovative insights to corresponding themes and incorporate forward-looking innovation directions into resource planning.

5.2. Prioritizing Innovation: The Value-Urgency Matrix

Based on the clarification of these five themes, we further analyzed the innovative comments and transformed the unstructured text into a more intuitive view. First, we continued to categorize and subdivide the primary themes into secondary themes using a method similar to the previous step. After the three experts provided their individual scores, we constructed a value–urgency matrix based on each category’s average score and the number of comments it contains (Figure 8). In the figure, bubble size corresponds to the number of comments under each secondary theme.

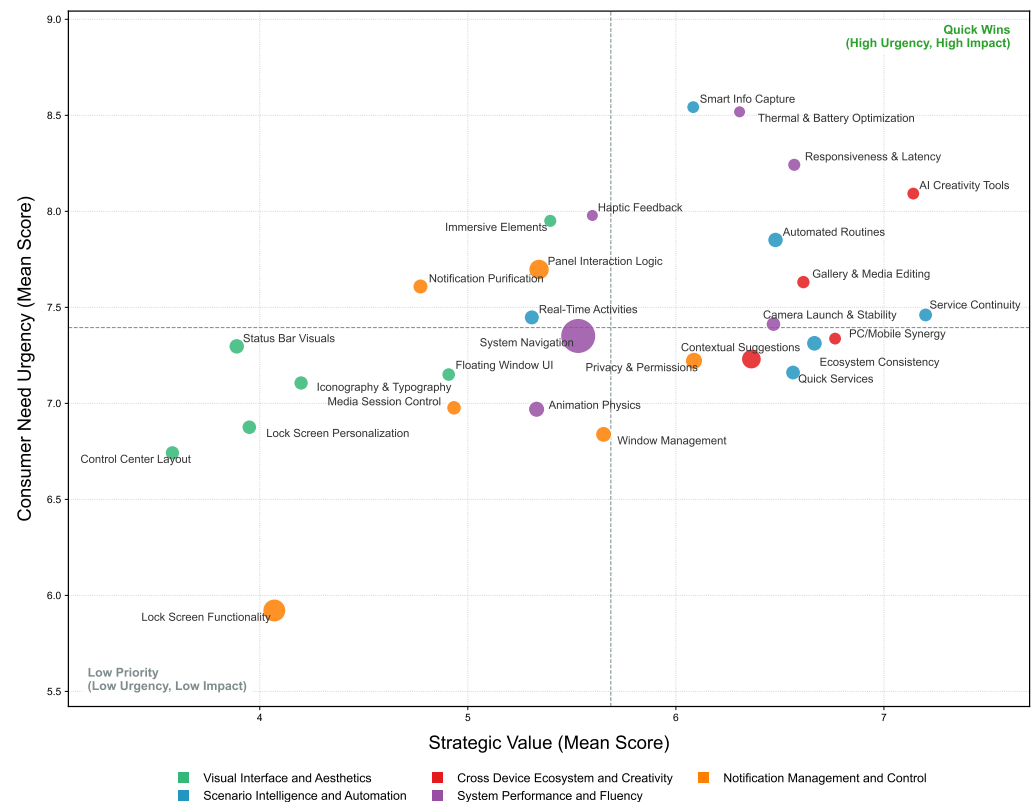


Figure 8. Value-Urgency Matrix for Innovation Prioritization.

This matrix clearly presents the development priorities of each theme, distinguishing between core issues and secondary improvement items. In the upper-right area, categories such as Service Continuity, Smart Info Capture, and AI Creativity Tools receive high scores on both dimensions. These functions are crucial for consumer experience and are likely to influence the system's competitive advantage. Among them, the prominent position of AI Creativity Tools further indicates consumers' potential demand for integrating more AI capabilities into the mobile operating system. In contrast, the lower-left area includes categories such as Lock Screen Functionality and Control Center Layout, which are relatively niche interface optimizations. Although some consumers mentioned these topics, their urgency and strategic value scores are relatively low.

This matrix not only clarifies the priority of secondary functional categories but also reflects the distribution of the five themes through color coding. For example, comments under the "Visual Interface and Aesthetics" theme are mostly concentrated in the lower right quadrant, indicating that this type of demand is usually viewed as an optimization item with low urgency and strategic value. By contrast, the "Scenario Intelligence and Automation" theme appears more often in the upper right quadrant, indicating that functions related to automation and contextual intelligence generally have higher consumer urgency and strategic potential.

Overall, the further analysis in this section indicates that the innovative insights identified by the framework can be transformed into structured planning inputs that are more conducive to discussion and prioritization. Specifically, the value-urgency matrix provides an intuitive way to present these insights, enabling the product team to make clearer comparisons and trade-offs across topics, thereby supporting product planning decisions [53].

6. Discussion

In the era of e-commerce and digital transformation, UGC has become an important source for brands to obtain innovative ideas [1,4]. Compared with more explicit text tasks such as sentiment and defect detection, innovative insights are often scarcer and expressed more implicitly [12,19], making it difficult for traditional models to identify them accurately. However, the value of a consumer innovation review may be much greater than that of a simple consumer complaint. Therefore, we propose a novel LLM-augmented Siamese framework and verify its effectiveness and cross-platform stability through multi-platform and multi-scenario experiments. The following will further elaborate on the theoretical contributions and managerial implications.

6.1. Academic Implications

Firstly, this paper advances UGC insights from “observing how consumers evaluate” to “observing what innovations consumers desire”, shifting the research focus from traditional emotion and defect analysis to a task that more directly serves product innovation—the identification of innovative insights. The systematic comparison results across different experimental setups in the Results section show that the framework in this paper performs better or more stably than traditional baselines and strong baselines in most scenarios, thereby providing transferable empirical evidence for identifying innovative insights from multi-platform UGC.

Secondly, this paper uses generative artificial intelligence not only to expand the sample, but also to conduct intent-preserving generation, thereby alleviating semantic ambiguity caused by information stickiness and the expression gap to some extent and making users’ implicit innovation intentions easier for the model to capture. Combined with the semantic consistency learning of the Siamese network, the model can focus more on the shared semantics behind different expressions, thereby improving the stable identification of implicit innovation intentions.

Thirdly, the cross-platform, multi-scenario results further show that although expression styles and community contexts differ significantly across platforms, the framework in this paper remains competitive in cross-platform tests such as LOPO. This suggests that, when dealing with diverse UGC, methods based on semantic representations and similarity learning are more reliable than paradigms that rely on fixed rules or shallow features.

6.2. Practical Implications

Firstly, this study provides a scalable technical path for brands to identify high-value innovative insights from a vast amount of UGC. Unlike traditional text mining methods mainly used for sentiment analysis and defect diagnosis, the framework in this paper is oriented toward identifying consumer innovation insights that are scarce and implicitly expressed. In most experimental settings, it performs better than both the baseline and the strong baseline, thereby providing a reusable technical solution for brands to identify innovative insights in multi-platform environments.

Secondly, this paper provides a clearer implementation path for practical decision making. After screening innovative comments, one can further conduct semantic clustering and topic summation of the identified insights and combine them with the “value–urgency” matrix for structured presentation. In this way, brands can organize scattered innovative insights into clear, hierarchical planning inputs that are comparable and convenient for discussion and prioritization, helping the product team allocate resources more systematically.

Finally, the experimental results across multiple platforms and scenarios show that although expression styles and community contexts vary across platforms, the framework in this paper remains competitive in cross-platform testing. This enables brands to more

consistently integrate innovative signals from different platforms, reduce the limitations caused by information silos and channel selection biases, and thereby support product planning and continuous iteration from a more comprehensive perspective.

7. Limitations and Future Research

Although this study has shown the effectiveness of the LLM-augmented Siamese framework for innovation insight mining under the challenges of data sparsity and information stickiness, several limitations remain, which also point to directions for future research.

First, the empirical validation in this study focuses on a single brand product, and all data are sourced from Chinese UGC platforms. Therefore, the transferability of the conclusions to other countries or language environments, as well as to different brands, still requires further verification. Future research can conduct comparisons across brands and languages to more systematically define the external validity and applicable boundaries of the framework.

Second, our analysis is based only on text data. However, modern UGC platforms are multimodal, and consumers often express complex ideas through images and videos. Relying only on text may miss information conveyed in visual content. A promising direction is to combine large models for multimodal processing [59] and integrate visual and textual information to better understand consumers' innovative ideas.

Third, although the LOPO multi-scenario analysis in this paper has verified the overall robustness of the framework, expression differences across UGC platforms still affect cross-platform generalization. Moreover, this study used only a single LLM for data augmentation. Future research can introduce adversarial training [60] and systematically compare different LLMs and enhancement strategies to further improve cross-platform recognition performance and generalization.

8. Conclusions

To address the challenge of identifying extremely rare but highly valuable consumer innovation insights from a vast amount of UGC, this study proposes a Siamese network framework augmented by large language models. By combining the text generation strengths of LLMs with the classification capability of a BERT-based architecture, our method effectively overcomes the limits of traditional machine learning and deep learning models in capturing subtle meaning and extreme class imbalance. Empirical results across multiple UGC datasets and experimental scenarios show that our method substantially outperforms benchmark models and can recall more innovative ideas with high precision. In summary, this study combines generative artificial intelligence with text classification, enabling brands to capture valuable innovation insights from user-generated content on social media and use these insights to improve products and services, thereby better meeting consumers' needs.

Author Contributions: J.W.: Writing—review & editing, Writing—original draft, Methodology, Data curation, Conceptualization. Q.W.: Writing—review & editing, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 72374188), and the Graduate Education Research Fund of USTC (Grant No. USTC-GERF24001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Additional Results

This appendix reports detailed single-platform and LOPO performance metrics.

Table A1. Single-platform performance metrics across four UGC platforms.

Model	Douyin				Xiaohongshu				Weibo				Bilibili			
	F1	PR-AUC	Precision	Recall	F1	PR-AUC	Precision	Recall	F1	PR-AUC	Precision	Recall	F1	PR-AUC	Precision	Recall
Model 1	0.316	0.167	0.333	0.300	0.278	0.212	0.556	0.185	0.485	0.604	0.446	0.532	0.468	0.321	0.379	0.611
Model 2	0.286	0.183	0.273	0.300	0.356	0.381	0.444	0.296	0.719	0.740	0.762	0.681	0.323	0.270	0.385	0.278
Model 3	0.667	0.793	0.636	0.700	0.667	0.719	0.667	0.667	0.857	0.931	0.886	0.830	0.714	0.880	1.000	0.556
Model 4	0.636	0.760	0.583	0.700	0.600	0.652	0.652	0.556	0.842	0.932	0.833	0.851	0.757	0.827	0.737	0.778
Model 5	0.667	0.812	0.636	0.700	0.679	0.749	0.692	0.667	0.882	0.942	0.891	0.872	0.765	0.816	0.812	0.722

Table A2. LOPO performance metrics across four UGC platforms.

Model	Douyin				Xiaohongshu				Weibo				Bilibili			
	F1	PR-AUC	Precision	Recall	F1	PR-AUC	Precision	Recall	F1	PR-AUC	Precision	Recall	F1	PR-AUC	Precision	Recall
Model 1	0.385	0.367	0.375	0.396	0.404	0.309	0.348	0.483	0.479	0.551	0.676	0.371	0.372	0.354	0.265	0.624
Model 2	0.469	0.432	0.535	0.418	0.529	0.545	0.614	0.466	0.660	0.705	0.780	0.573	0.447	0.433	0.483	0.416
Model 3	0.817	0.876	0.885	0.758	0.816	0.890	0.850	0.784	0.909	0.957	0.937	0.882	0.827	0.896	0.804	0.851
Model 4	0.810	0.858	0.883	0.747	0.820	0.890	0.881	0.767	0.801	0.953	0.981	0.677	0.758	0.902	0.655	0.901
Model 5	0.810	0.881	0.883	0.747	0.819	0.893	0.915	0.741	0.892	0.964	0.935	0.852	0.762	0.896	0.659	0.901

References

- Bayus, B.L. Crowdsourcing new product ideas over time: An analysis of the dell ideastorm community. *Manag. Sci.* **2013**, *59*, 226–244. [\[CrossRef\]](#)
- Nambisan, S.; Lyytinen, K.; Majchrzak, A.; Song, M. Digital innovation management. *MIS Q.* **2017**, *41*, 223–238. [\[CrossRef\]](#)
- Randhawa, K.; Wilden, R.; Hohberger, J. A bibliometric review of open innovation: Setting a research agenda. *J. Prod. Innov. Manag.* **2016**, *33*, 750–772. [\[CrossRef\]](#)
- Nasrabadi, M.A.; Beauregard, Y.; Ekhlassi, A. The implication of user-generated content in new product development process: A systematic literature review and future research agenda. *Technol. Forecast. Soc. Change* **2024**, *206*, 123551. [\[CrossRef\]](#)
- Vargo, S.L.; Lusch, R.F. Institutions and axioms: An extension and update of service-dominant logic. *J. Acad. Mark. Sci.* **2016**, *44*, 5–23. [\[CrossRef\]](#)
- Zheng, X.; Cheung, C.M.; Lee, M.K.; Liang, L. Building brand loyalty through user engagement in online brand communities in social networking sites. *Inf. Technol. People* **2015**, *28*, 90–106. [\[CrossRef\]](#)
- Kilumile, J.W.; Zuo, L. The nexus of influencers and purchase intention: Does consumer brand co-creation behavior matter? *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 3088–3101. [\[CrossRef\]](#)
- Zhang, C.; Xu, Z. Gaining insights for service improvement through unstructured text from online reviews. *J. Retail. Consum. Serv.* **2024**, *80*, 103898. [\[CrossRef\]](#)
- Chen, H.; Chiang, R.H.; Storey, V.C. Business intelligence and analytics: From big data to big impact. *Mis Q.* **2012**, *36*, 1165–1188. [\[CrossRef\]](#)
- Erevelles, S.; Fukawa, N.; Swayne, L. Big data consumer analytics and the transformation of marketing. *J. Bus. Res.* **2016**, *69*, 897–904. [\[CrossRef\]](#)
- Von Hippel, E. “Sticky information” and the locus of problem solving: Implications for innovation. *Manag. Sci.* **1994**, *40*, 429–439. [\[CrossRef\]](#)
- Timoshenko, A.; Hauser, J.R. Identifying customer needs from user-generated content. *Mark. Sci.* **2019**, *38*, 1–20. [\[CrossRef\]](#)
- Wang, L.; Che, G.; Hu, J.; Chen, L. Online review helpfulness and information overload: The roles of text, image, and video elements. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1243–1266. [\[CrossRef\]](#)
- Büschken, J.; Allenby, G.M. Sentence-based text analysis for customer reviews. *Mark. Sci.* **2016**, *35*, 953–975. [\[CrossRef\]](#)
- Abrahams, A.S.; Fan, W.; Wang, G.A.; Zhang, Z.; Jiao, J. An integrated text analytic framework for product defect discovery. *Prod. Oper. Manag.* **2015**, *24*, 975–990. [\[CrossRef\]](#)
- Mustak, M.; Hallikainen, H.; Laukkanen, T.; Plé, L.; Hollebeek, L.D.; Aleem, M. Using machine learning to develop customer insights from user-generated content. *J. Retail. Consum. Serv.* **2024**, *81*, 104034. [\[CrossRef\]](#)
- Shen, Z.; Zhao, C.; Li, Y. Customer requirements analysis and product service improvement framework using multi-source user-generated content and dual importance–performance analysis: A case study of fresh e-commerce. *J. Theor. Appl. Electron. Commer. Res.* **2026**, *21*, 19. [\[CrossRef\]](#)

18. Cui, X.; Zhu, Z.; Liu, L.; Zhou, Q.; Liu, Q. Anomaly detection in consumer review analytics for idea generation in product innovation: Comparing machine learning and deep learning techniques. *Technovation* **2024**, *134*, 103028. [\[CrossRef\]](#)
19. Zhang, M.; Fan, B.; Zhang, N.; Wang, W.; Fan, W. Mining product innovation ideas from online reviews. *Inf. Process. Manag.* **2021**, *58*, 102389. [\[CrossRef\]](#)
20. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084. [\[CrossRef\]](#)
21. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
22. Dai, H.; Liu, Z.; Liao, W.; Huang, X.; Cao, Y.; Wu, Z.; Zhao, L.; Xu, S.; Zeng, F.; Liu, W.; et al. Auggpt: Leveraging chatgpt for text data augmentation. *IEEE Trans. Big Data* **2025**, *11*, 907–918. [\[CrossRef\]](#)
23. Ding, B.; Qin, C.; Zhao, R.; Luo, T.; Li, X.; Chen, G.; Xia, W.; Hu, J.; Luu, A.T.; Joty, S. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv* **2024**, arXiv:2403.02990.
24. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.
25. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol. (TIST)* **2022**, *13*, 31. [\[CrossRef\]](#)
26. Berger, J.; Humphreys, A.; Ludwig, S.; Moe, W.W.; Netzer, O.; Schweidel, D.A. Uniting the tribes: Using text for marketing insight. *J. Mark.* **2020**, *84*, 1–25. [\[CrossRef\]](#)
27. Zhu, Q.; Wang, Y.; Xu, X.; Sarkis, J. How loud is consumer voice in product deletion decisions? retail analytic insights. *J. Retail. Consum. Serv.* **2025**, *82*, 104110. [\[CrossRef\]](#)
28. Archak, N.; Ghose, A.; Ipeirotis, P.G. Deriving the pricing power of product features by mining consumer reviews. *Manag. Sci.* **2011**, *57*, 1485–1509. [\[CrossRef\]](#)
29. Tirunillai, S.; Tellis, G.J. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *J. Mark. Res.* **2014**, *51*, 463–479. [\[CrossRef\]](#)
30. Maalej, W.; Nabil, H. Bug report, feature request, or simply praise? On automatically classifying app reviews. In *2015 IEEE 23rd International Requirements Engineering Conference (RE)*; IEEE: Piscataway, NJ, USA, 2015; pp. 116–125.
31. Govindarajan, V.S.; Chen, B.; Warholc, R.; Erk, K.; Li, J.J. Help! Need advice on identifying advice. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 8–12 November 2020; pp. 5295–5306.
32. Kühl, N.; Mühlthaler, M.; Goutier, M. Supporting customer-oriented marketing with artificial intelligence: Automatically quantifying customer needs from social media. *Electron. Mark.* **2020**, *30*, 351–367. [\[CrossRef\]](#)
33. Di Marco, N.; Loru, E.; Bonetti, A.; Serra, A.O.G.; Cinelli, M.; Quattrociochi, W. Patterns of linguistic simplification on social media platforms over time. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2412105121. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A survey of data augmentation approaches for nlp. *arXiv* **2021**, arXiv:2105.03075. [\[CrossRef\]](#)
35. Wei, J.; Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv* **2019**, arXiv:1901.11196. [\[CrossRef\]](#)
36. Liu, Q.; Du, Q.; Hong, Y.; Fan, W.; Wu, S. User idea implementation in open innovation communities: Evidence from a new product development crowdsourcing community. *Inf. Syst. J.* **2020**, *30*, 899–927. [\[CrossRef\]](#)
37. Zhang, S.; Pan, S.L.; Ouyang, T.H. Building social translucence in a crowdsourcing process: A case study of miui.com. *Inf. Manag.* **2020**, *57*, 103172. [\[CrossRef\]](#)
38. Uysal, A.K.; Gunal, S. The impact of preprocessing on text classification. *Inf. Process. Manag.* **2014**, *50*, 104–112. [\[CrossRef\]](#)
39. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1971**, *76*, 378. [\[CrossRef\]](#)
40. Lin, Y.C.; Chen, S.A.; Liu, J.J.; Lin, C.J. Linear classifier: An often-forgotten baseline for text classification. *arXiv* **2023**, arXiv:2306.07111. [\[CrossRef\]](#)
41. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781. [\[CrossRef\]](#)
43. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with whole word masking for Chinese bert. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [\[CrossRef\]](#)
44. Bouschery, S.G.; Blazevic, V.; Piller, F.T. Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *J. Prod. Innov. Manag.* **2023**, *40*, 139–153. [\[CrossRef\]](#)
45. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. Deepseek-v3 technical report. *arXiv* **2024**, arXiv:2412.19437.
46. Li, B.; Hou, Y.; Che, W. Data augmentation approaches in natural language processing: A survey. *Ai Open* **2022**, *3*, 71–90. [\[CrossRef\]](#)
47. Tenney, I.; Das, D.; Pavlick, E. Bert rediscovers the classical nlp pipeline. *arXiv* **2019**, arXiv:1905.05950. [\[CrossRef\]](#)

48. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
49. Izmailov, P.; Podoprikin, D.; Garipov, T.; Vetrov, D.; Wilson, A.G. Averaging weights leads to wider optima and better generalization. *arXiv* **2018**, arXiv:1803.05407.
50. Cohen, M.A.; Eliasberg, J.; Ho, T.H. New product development: The performance and time-to-market tradeoff. *Manag. Sci.* **1996**, *42*, 173–186. [[CrossRef](#)]
51. Si, H.; Kavadias, S.; Loch, C. Managing innovation portfolios: From project selection to portfolio design. *Prod. Oper. Manag.* **2022**, *31*, 4572–4588. [[CrossRef](#)]
52. Svensson, R.B.; Torkar, R. Not all requirements prioritization criteria are equal at all times: A quantitative analysis. *J. Syst. Softw.* **2024**, *209*, 111909. [[CrossRef](#)]
53. Nagji, B.; Tuff, G. Managing your innovation portfolio. *Harv. Bus. Rev.* **2012**, *90*, 66–74.
54. Valverde-Albacete, F.J.; Peláez-Moreno, C. 100% classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE* **2014**, *9*, e84217. [[CrossRef](#)] [[PubMed](#)]
55. Davis, J.; Goadrich, M. The relationship between precision-recall and roc curves. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 233–240.
56. Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Qin, T.; Lu, W.; Chen, Y.; Zeng, W.; Yu, P.S. Generalizing to unseen domains: A survey on domain generalization. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8052–8072. [[CrossRef](#)]
57. Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [[CrossRef](#)]
58. Maaten, L.v.d.; Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
59. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *Natl. Sci. Rev.* **2024**, *11*, nwae403. [[CrossRef](#)]
60. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.