AUDCRAFT: UNDERSTANDING THE AUDIENCE-ORIENTED CREATION GAP BETWEEN HUMAN CREATORS AND LLMS WITH REAL-WORLD PLATFORM DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

As low-cost LLM-generated content enters online platforms at scale, understanding how LLMs create for different audiences is increasingly important for assessing their potential impact on content ecosystems. Yet existing evaluations largely treat LLM-generated content as isolated textual artifacts, focusing on properties such as creativity and detectability while overlooking a key aspect of real-world content creation: creators target specific audiences and adapt content to their needs. To address this gap, we introduce AUDCRAFT, an observational corpus of real-world human-created content spanning diverse audiences, and evaluate LLMs from two complementary dimensions of audience-oriented creation: *audience orientation*, which characterizes whom LLMs create for, and *audience design*, which characterizes how they adapt their expression to different audiences. Our analysis reveals systematic human-LLM gaps in both dimensions. For audience orientation, LLMs disproportionately concentrate their generations on a narrow subset of audience segments, a pattern that persists across models and prompting strategies. A Tullock contest-based simulation further shows that such concentrated orientation can be amplified through competition for limited audience attention, resulting in increasingly uneven content supply and intensified competition with human creators. For audience design, LLMs exhibit rich and audience-responsive expressions, but this flexibility does not necessarily translate into human-like adaptation: their changes are selective, often misaligned with human patterns, and less audience-specific in vocabulary. Code and data are available at `https://anonymous.4open.science/r/AudCraft-54E4`.

1 INTRODUCTION

By enabling fluent text generation at substantially lower marginal cost, large language models (LLMs) make it possible to produce content at a scale that was previously difficult to sustain through human labor alone (Tiwari, 2023; Li et al., 2026; Wei & Tyson, 2024). This shift has raised growing concerns about how scalable AI-generated content (AIGC) may reshape the content supply landscape of real-world content platforms (e.g., Xiaohongshu, TikTok) (Lam et al., 2020; Ren, 2024; Abdollahpouri et al., 2020), alter competition among creators (Li, 2024; Kim et al., 2026), and reallocate limited user attention (Yao et al., 2024a; Katsamakas & Sánchez-Cartas, 2024). Understanding how the growing presence of LLM-generated content may transform content ecosystems is therefore increasingly important.

Existing studies often treat LLM-generated content as isolated textual artifacts and compare it directly with human-written text. For example, prior work has examined differences between LLM-generated and human-written content in terms of rhetorical patterns (Reinhart et al., 2025), writing styles (Rallapalli et al., 2026), and stylometric characteristics (Tripto et al., 2025). Another line of work investigates whether LLMs can imitate human creators, such as generating user-conditioned social media posts (Davoudi et al., 2026), mimicking literary and political writing styles (Alsadhan, 2026), and producing human-like narratives (Zhao et al., 2023). More recently, researchers have

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

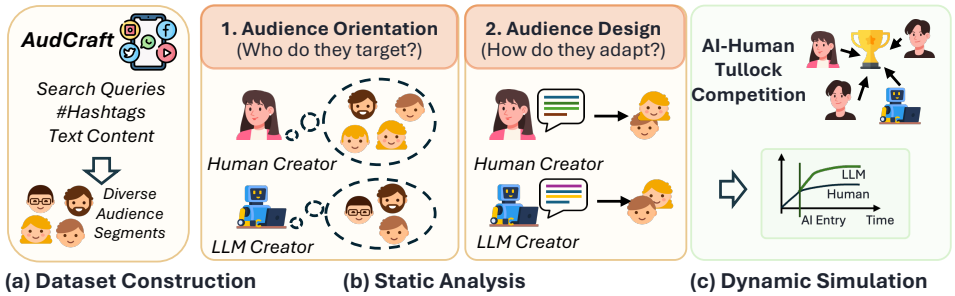


Figure 1: Workflow of the proposed audience-oriented evaluation framework.

begun to examine the ecosystem-level effects of AIGC, showing, for example, that AIGC can affect retrieval systems and potentially alter content exposure dynamics (Dai et al., 2024).

However, these perspectives provide limited insight into audience-oriented content creation on real-world platforms (Bell, 1984). Sociolinguistics research on audience design suggests that content creation on online platforms is a communicative act directed toward intended audiences (Bell, 1984; Litt & Hargittai, 2016). Human creators often adjust their creation strategies according to the audiences they intend to reach (Bell, 1984). For instance, a travel guide targeting trip planners may adopt a well-structured format with concrete details (e.g., checklists), whereas lifestyle-sharing content may use a more casual tone and affective expression to resonate with audiences. Such audience-oriented strategies are especially consequential on content platforms, where creators targeting the same audience often compete for the same pool of user traffic (Abdollahpouri et al., 2020; Yao et al., 2024b).

Understanding how LLM-generated content may reshape content ecosystems therefore requires moving beyond text-level properties to examine audience-oriented creation. We study this problem from two complementary perspectives: (1) **RQ1: Audience Orientation.** Whom do LLMs create for, and how does their audience orientation differ from that of human creators? (2) **RQ2: Audience Design.** How do LLMs adapt their content to different audiences, and how does such adaptation differ from that of human creators?

To answer these questions, we introduce AUDCRAFT, a large-scale dataset collected from a Chinese content platform, providing a real-world reference for comparing human-created and LLM-generated content across diverse audience domains and intents. Drawing on prior studies of audience analysis and audience design (McQuail, 1997; Bell, 1984), we systematically evaluate LLMs along these two dimensions.

For *Audience Orientation* (RQ1), we find that LLMs concentrate their generations on a narrower range of audience segments than human creators, a pattern that persists across models and prompting strategies. We further construct a repeated competition environment based on the Tullock contest model (Skaperdas, 1996b; Tullock et al., 1980) to examine its potential ecosystem impact. Our simulation suggests that concentrated LLM audience orientation can be amplified through competition for limited audience attention, leading to increasingly uneven content supply and intensified competition with human creators. For *Audience Design* (RQ2), we find that LLMs exhibit rich and audience-responsive expressions, showing substantial adaptation in several stylistic dimensions. However, such flexibility does not necessarily translate into human-like audience design: their adaptations are highly selective and often weakly aligned with how human creators differentiate their expressions across audiences. This gap is particularly evident in audience-specific language, where LLMs systematically underuse the distinctive vocabulary observed in human content.

Altogether, our work introduces an audience-oriented framework for evaluating LLMs in real-world content creation, integrating AUDCRAFT, systematic analyses of audience orientation and audience design, and an exploratory competition-based simulation. By examining both *whom* LLMs create for and *how* they adapt content to different audiences, our framework provides a useful resource for diagnosing human–LLM gaps that are overlooked by conventional text-centric evaluations. Our findings further reveal how concentrated audience orientation may propagate through platform

Table 1: Representative real-world HGC cases across audience segments.

Audience Intent	Representative Case
Food Enthusiasts	**Title.** *Luckin Coffee Through a Fat-Loss Lens, Part V* / 减脂人眼里的瑞幸咖啡｜第五期 **Excerpt.** *A low-calorie Luckin drink guide, from a 150-calorie Fair Isle Latte to a 255-calorie Green Apple Velvet Latte, with ordering tips for reducing sugar.* / 瑞幸低卡饮品攻略：从150大卡的费尔岛拿铁到255大卡的青苹果丝绒拿铁，并提供减糖点单建议。 **Hashtags.** *Luckin low-calorie drinks; calorie guide; coffee review* / #瑞幸低卡饮品攻略；#瑞幸热量攻略；#咖啡测评 **Query.** *Luckin coffee calories (1); low-calorie recommendations (1)* / 瑞幸咖啡热量 (1)；瑞幸低卡推荐 (1)
Travel Explorers	**Title.** *Eight Gorgeous Cities for Fall and Winter Trips* / 适合秋冬去的8大城市，每一个都超美 **Excerpt.** *Eight cities for fall and winter trips; in Chongqing, the light rail covers most attractions, costs 2–7 RMB, and avoids traffic.* / 整理了8个适合秋冬旅行的城市；重庆推荐轻轨出行，覆盖大多数景点，票价2–7元且不堵车。 **Hashtags.** *Chongqing travel; fall/winter destinations; travel guide* / #重庆旅游；#秋冬去哪玩；#游玩攻略 **Query.** *winter travel destinations (1); recommended travel cities (1)* / 冬天适合去旅游 (1)；旅游推荐城市 (1)
Wellness Seekers	**Title.** *If Your Nose Is Blocked, Try This for Instant Relief* / 鼻塞的宝子们都来试试，鼻子一秒通畅 **Excerpt.** *Six quick ways to relieve nasal congestion and breathe more comfortably, especially when a blocked nose makes sleeping difficult.* / 分享6个快速缓解鼻塞的方法，帮助改善鼻塞带来的呼吸不畅和夜间睡眠困难。 **Hashtags.** *nasal congestion; blocked-nose remedies; health and wellness* / #鼻塞；#鼻塞怎么办；#健康养生 **Query.** *clear a blocked nose (1); fastest congestion relief (1)* / 鼻子堵塞通气 (1)；感冒鼻塞最快解决办法 (1)

competition, providing insights for guiding the development and governance of LLM-generated content toward more diverse and healthy content ecosystems.

2 AUDCRAFT: REAL-WORLD OPEN-ENDED CONTENTS WITH DIVERSE AUDIENCES

Understanding how LLM-generated content may reshape content ecosystems requires knowing what kinds of audiences real-world human content is created for. To provide a real-world reference for audience-oriented content creation, we construct AUDCRAFT, a corpus of real-world, open-ended human-written content with diverse audiences, based on *Qilin* (Chen et al., 2025), a large-scale open-source Chinese UGC dataset[1].

Mining in-the-wild human-written posts. We construct AUDCRAFT, a dataset of real-world, open-ended human-written content with diverse audience taxonomies, by filtering and refining posts from *Qilin*. This corpus serves as the real-world human reference for our comparison with AIGC. We retain Chinese posts with sufficient textual content and remove duplicated or corrupted samples. To reduce potential contamination from AIGC, we further apply a widely-used AIGC detector (Bao et al., 2024) to exclude posts likely generated by AI. The resulting corpus contains 1,590,249 posts and serves as the human-written reference for our comparison. Full statistics for each filtering step are reported in Appendix B.

Constructing an audience taxonomy with content and engagement signal. Directly inferring the audience intent from content text alone is challenging because it is often implicit and not fully expressed in the text (Bell, 1984). However, previous works shows that creator-provided affordances (e.g., hashtags (Pavalanathan & Eisenstein, 2015)) and user interactions (e.g., search query (Xiong et al., 2013)) provide observable cues about how creators position their posts toward particular audience segments. Hence, we ground taxonomy construction in complementary supply- and demand-side evidence: For each content item x_n, we consider its content text and hashtags as supply-side signals, together with a set of engagement-positive search queries $\mathcal{Q}_n$ as demand-side signals. Specifically, $\mathcal{Q}_n$ contains queries through which users were exposed to x_n and subsequently engaged with it (e.g., click, like, comment, or collect). These queries provide behavioral evidence of the information needs of users who actually interacted with the content. The example is illustrated in Table 1.

To characterize the diverse audiences of real-world content, we develop a fine-grained audience taxonomy along two complementary dimensions, (1) *audience domain*, which captures the broad sphere of user interest, and (2) *audience intent*, which captures the underlying information need, following prior studies on audience segmentation (Smith, 1956).

[1] https://huggingface.co/datasets/THUIR/Qilin

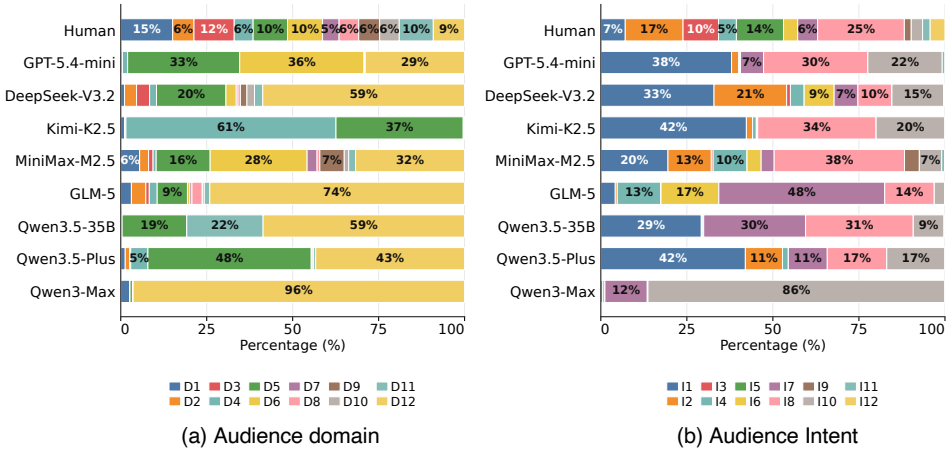
We adopt a semi-automatic process to construct the taxonomy. Starting from a stratified sample of HGCs, human annotators jointly examine the content, hashtags, and associated queries to identify recurring audience needs and derive tentative domain and intent labels. We iteratively refine these seed labels by merging redundant categories, splitting overly broad ones, removing topic-specific labels, and clarifying category boundaries. We then use `gpt-5.5` to scale the analysis to a larger stratified sample by assigning existing labels and flagging potentially uncovered audience types. Human annotators review these flagged cases and make the final decisions on taxonomy updates. The resulting fixed taxonomy contains 12 audience domains and 12 audience intents. Details of the taxonomy construction and annotation procedures are provided in Appendix B.2.

3 Audience Orientation Evaluation (RQ1)

We first investigate *Audience Orientation*, focusing on which audiences LLMs naturally target during open-ended content creation. Using AudCraft as a real-world human reference, we examine whether LLMs exhibit audience-targeting patterns distinct from human creators, whether these patterns persist across different generation settings, and how they may evolve under competition for limited audience attention.

3.1 Evaluation Setups

Using a random subset of 100 human-created posts from AudCraft and 100 posts independently generated by each LLM without specifying any target audience or topic, we compare their audience orientations. Each post is manually annotated with an audience domain and intent by a pool of 10 human annotators, who receive only the title, body, and fixed taxonomy definitions and are blinded to the content source and model identity. We compare the resulting audience distributions to assess whether LLM-generated content concentrates on a narrower range of audiences than human-created content. Detailed generation prompts, annotation procedures, and complete results are provided in Appendix D.2.



(a) Audience domain (b) Audience Intent

Figure 2: Audience orientation distributions of human-created and LLM-generated content across (a) audience domains and (b) audience intents. LLM-generated content is more concentrated on a smaller set of domains and intents, indicating more homogenized audience orientation. Full label definitions are provided in Table 3.

3.2 Audience Orientation Analysis

Audience orientation distribution of human and LLMs. Figure 2 compares the audience orientation distributions of human creators and different LLMs. The domain-level distribution in Figure 2 (a) shows that HGC is relatively balanced across audience domains, with no single domain dominating the corpus. In contrast, LLMs often concentrate on a small number of domains. For example, `GPT-5.4-mini` mainly targets *Food Enthusiasts*, *Travel Explorers*, and *Self-Improvers*, while

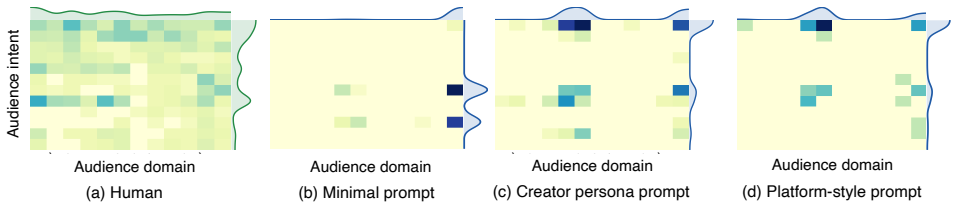


Figure 3: Audience orientation of HGC and `gpt-5.4-mini` generations under different prompting strategies. All LLM outputs are generated with temperature $\tau = 1$. Darker colors indicate higher concentration in the corresponding audience segment.

`Qwen3-Max` assigns 96% of its generations to *Self-Improvers*. Such concentration indicates that LLMs do not reproduce the broad audience coverage of human creators, but instead favor a narrower set of platform audiences.

Figure 2(b) shows a similar concentration pattern at the intent level. While HGC covers a broad range of audience intents, LLM-generated content tends to concentrate on a few recurring intents, such as *Learning from Experience*, *Validating a Choice*, and *Engaging with a Story*. For example, `Kimi-K2.5` allocates over 96% of its generations to these three intents, and `Qwen3-Max` assigns 86% of its outputs to *Engaging with a Story*. This suggests that LLMs exhibit a more concentrated inferred distribution over audience-intent categories than real-world creators.

Homogenized audience orientation across prompting strategies. As shown in Figure 3, the concentrated audience orientation of LLM-generated content persists across different prompting strategies. While HGC broadly covers diverse domain–intent segments, LLM generations consistently cluster around a small subset of audiences, leaving many segments rarely targeted. Notably, providing richer prompting contexts, such as creator personas or platform-specific instructions, does not substantially alleviate this concentration. This consistency suggests that the observed audience-orientation gap is not specific to a particular prompting strategy, but reflects a persistent tendency of LLMs toward a narrower set of audiences.

3.3 Dynamic Impact of Audience Orientation

To examine how the audience-orientation gap evolves under competition for limited attention, we build a Tullock contest-based simulation (Tullock et al., 1980), initialized with human and LLM audience-orientation distributions estimated from AudCraft and `gpt-5.4-mini`, respectively. The simulation workflow is detailed in Algorithm 1.

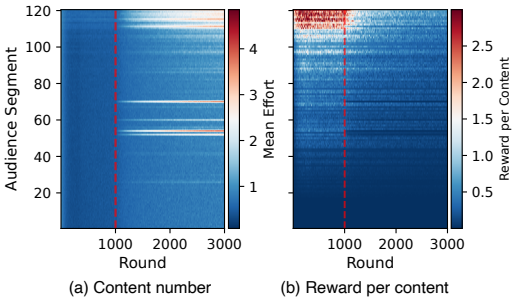


Figure 4: Changes in (a) content supply and (b) average reward across audience segments, sorted from the least (1) to the most (120) popular.

Environment Construction. We model the platform as a set of audience segments with finite attention pools, where creators targeting the same segment compete for exposure (Yao et al., 2024b; Lin et al., 2024; Ye et al., 2024; Xu et al., 2024). Creator rewards depend on their own effort relative to the aggregate effort within the targeted segment. Before LLM entry, creators follow the human audience-orientation distribution; after entry, a subset follows the LLM distribution. We track the resulting changes in content supply, rewards, and audience coverage. Full simulation settings are provided in Appendix E.

Simulation Results and Insights As shown in Figure 4(a), after AI entry at round 1,000, content supply becomes increasingly concentrated in popular and LLM-preferred audience segments, while many other segments remain weakly covered. This increased supply intensifies competition for limited attention: Figure 4(b) shows declining average content rewards, particularly in segments favored by LLM creators.

Figure 5 further shows that AI-assisted creators capture an increasing share of platform rewards as AI creation costs decrease, accompanied by growing reward losses for human creators. Together, these results suggest that concentrated LLM audience orientation can be amplified through competition, leading to more uneven content supply and stronger competition with human creators. Additional results are provided in Appendix E.3.

4 AUDIENCE DESIGN EVALUATION (RQ2)

We next examine *Audience Design*, the phenomenon whereby creators adapt their communicative strategies to their intended audiences (Bell, 1984). Using AUDCRAFT as a real-world human reference, we compare how LLMs and human creators adjust their linguistic and stylistic choices across different audience segments.

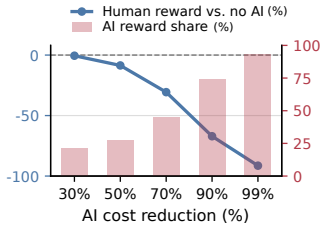


Figure 5: Changes in human and AI creator rewards after AI entry under different AI adoption rates and creation costs.

4.1 EVALUATION SETUPS

We examine whether LLMs reproduce *audience design*, the phenomenon whereby human creators adapt their framing, wording, and style to their intended audiences (Bell, 1984). We represent each audience segment as $a = (d, i)$, where $d \in \mathcal{D}$ denotes the audience domain and $i \in \mathcal{I}$ denotes the audience intent. We exclude extremely rare domain–intent combinations and retain 55 audience segments $\mathcal{A}$ that are sufficiently represented in AUDCRAFT. For each audience segment, we sample 100 human-written posts and generate 100 posts from each of four LLMs (i.e., GPT-5.4 mini, DeepSeek V3.2, Qwen3.5 Plus, and Kimi K2.5) using prompts that specify the corresponding audience domain, intent, and their short definitions.

We characterize each post using 14 expression metrics spanning length and readability, lexical diversity, style, and discourse, and audience targeting. Let $x_{s,a,k,i}$ denote the value of metric k for the i-th post from source $s \in \{\text{Human}, \text{LLM}\}$ targeting audience segment a, where s denotes either human content or one of the four LLMs. We summarize each source–audience pair by $\bar{x}^s_{a,k} = \frac{1}{100}\sum_{i=1}^{100} x^s_{a,k,i}$. The resulting audience-level profiles $\mathbf{x}^s_k = \bar{x}^s_{a,k}, a \in \mathcal{A}$ characterize how each expression metric varies across the 55 audience segments for humans and LLMs.

4.1.1 STYLOMETRIC ANALYSIS

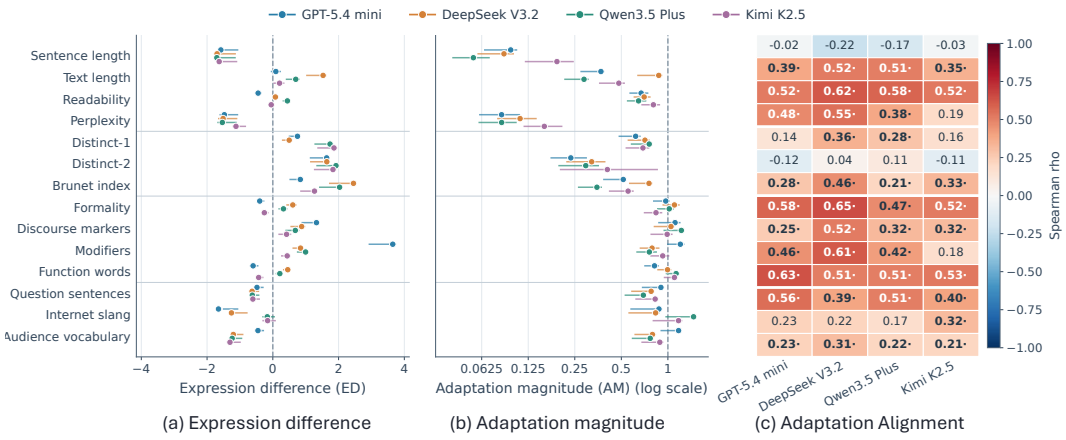
We use these profiles to examine audience design from three complementary perspectives: whether LLMs match human expression levels (*level*), whether they vary their expressions across audiences to a similar extent (*magnitude*), and whether these variations occur for the same audiences as in human content (*alignment*).

Overall Expression Difference (ED). We first examine whether LLM-generated content matches the overall expression levels observed in human content. For each metric k, we quantify the shift of model L from the human reference as

$$\text{ED}^{\text{LLM}}_k = \frac{\text{mean}_a(\bar{x}^{\text{LLM}}_{a,k}) - \text{mean}_a(\bar{x}^{\text{Human}}_{a,k})}{\text{SD}_a(\bar{x}^{\text{Human}}_{a,k})}, \tag{1}$$

where the difference is normalized by the cross-audience standard deviation of human content $\text{SD}_a(\bar{x}^{\text{Human}}_{a,k})$. Thus, ED $= 0$ indicates matching average levels, while its sign indicates the direction of the shift and its absolute magnitude reflects the extent of deviation in units of human cross-audience variation.

Figure 6 (a) reveals substantial and systematic expression-level shifts across models. Notably, LLM-generated content tends to exhibit higher lexical diversity than human content, yet substantially lower audience-targeting scores, suggesting that lexical richness does not necessarily translate into stronger audience-specific expression. More broadly, the persistent shifts across multiple metrics indicate that LLMs retain characteristic model-specific expression baselines even when explicitly conditioned on



(a) Expression difference (b) Adaptation magnitude (c) Adaptation Alignment

Figure 6: Audience adaptation of LLM-generated content compared with HGC. (a) Overall expression difference: standardized LLM–human difference, where 0 indicates no shift. (b) Audience adaptation magnitude: LLM-to-human ratio of cross-audience variation, where 1 indicates human-level variation. (c) Audience adaptation alignment: Spearman correlation between LLM and human audience profiles. Intervals are 95% CIs from 2,000 bootstrap resamples.

different target audiences. Whether and to what extent they can move beyond these baselines to adapt their expressions across audiences is examined next.

Audience Adaptation Magnitude (AM). Expression-level shifts alone do not reveal whether LLMs adjust their expressions across different audiences. We therefore examine the magnitude of cross-audience variation. For each source s and metric k, we compute the standard deviation of audience-level means, $\sigma_k^s = \mathrm{SD}_a(\bar{x}_{a,k}^s)$, and quantify the relative adaptation magnitude of model L as $\mathrm{AM}_k^{\mathrm{LLM}} = \frac{\sigma_k^{\mathrm{LLM}}}{\sigma_k^{\mathrm{Human}}}$. A ratio of 1 indicates human-level cross-audience variation, while a ratio below (above) 1 indicates that the model varies less (more) across audiences than humans.

Figure 6(b) shows that LLMs do adjust their expressions across audiences, but such adaptation is highly selective across linguistic dimensions. Sentence length, perplexity, and lexical-diversity metrics exhibit substantially weaker cross-audience variation than in human content, suggesting that these characteristics remain relatively stable despite changes in the target audience. In contrast, formality, discourse markers, function words, question usage, and Internet slang show stronger audience-dependent variation, in some cases approaching or exceeding human levels. Overall, however, LLMs tend to exhibit less cross-audience variation than humans across most expression metrics. These results suggest that LLMs are responsive to audience information, but their adaptation is concentrated in a limited subset of expression dimensions rather than uniformly reflected across their linguistic behavior.

Audience Adaptation Alignment (AA). Cross-audience variation alone does not indicate whether LLMs adapt their expressions to audiences in the same way as human creators. A model may vary substantially across audiences while making these changes for different audiences than humans do. We therefore measure *adaptation alignment* using the Spearman correlation (SPR) between human and LLM audience-level profiles: $\mathrm{AA}_k^{\mathrm{LLM}} = \mathrm{SPR}\left(\left\{\bar{x}_{a,k}^{\mathrm{Human}}\right\}_{a=1}^{55}, \left\{\bar{x}_{a,k}^{\mathrm{LLM}}\right\}_{a=1}^{55}\right)$. This metric captures whether humans and LLMs rank the 55 audience segments similarly with respect to a given expression feature, with a higher positive correlation indicating stronger human–LLM alignment in audience-specific adaptation.

Figure 6 (c) shows that LLMs exhibit only partial alignment with human audience-specific adaptation patterns, with all observed correlations remaining below 0.7. For dimensions such as formality, readability, function-word usage, and question usage, LLMs exhibit moderate positive correlations with human patterns, suggesting that they can partially identify which audiences call for stronger or weaker use of these expressions. In contrast, sentence length and Distinct-2 show little consistent alignment, indicating that their cross-audience variations do not reliably follow human audience-specific patterns.

Importantly, adaptation magnitude and adaptation alignment capture distinct aspects of audience design. A model may exhibit substantial cross-audience variation without varying in a human-aligned manner. For example, Qwen3.5 Plus reaches an Internet Slang Ratio adaptation magnitude of AM $= 1.47$, exceeding the human level, yet its alignment with the human audience ordering remains weak (AA $= 0.17$). Thus, greater cross-audience variation does not necessarily imply better audience adaptation: LLMs may change their expressions across audiences, but not necessarily for the same audiences or in the same direction as human creators.

4.1.2 Audience-specific Vocabulary Analysis.

Beyond general expression patterns, we examine whether LLMs use vocabulary that is specifically associated with their target audiences. For each audience segment a, we construct an audience-specific vocabulary from human-written posts outside the evaluation set, selecting the 30 most discriminative words for that audience. For a post i generated by source s, we define its audience vocabulary score as $V^s_{a,i} = \frac{\#\text{tokens matching the audience-specific vocabulary of } a}{\#\text{valid tokens}}$ and average the score over the 100 posts for each audience: $\bar{V}^s_a = \frac{1}{100}\sum_{i=1}^{100} V^s_{a,i}$.

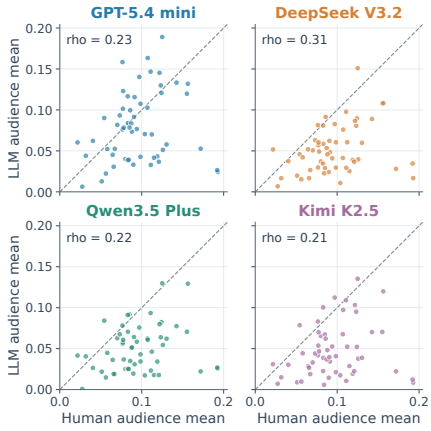


Figure 7: Audience-specific vocabulary usage of LLMs compared with HGC across 55 audience segments. The dashed line indicates $y = x$.

Figure 7 compares the resulting scores between humans and each LLM across the 55 audience segments. Each point represents one audience, with the human score on the x-axis and the corresponding LLM score on the y-axis; the diagonal indicates perfect agreement. We further report Spearman's ρ to measure whether LLMs reproduce the relative audience differences observed in human content.

As shown in Figure 7, most audience segments fall below the diagonal, indicating that LLM-generated content generally contains fewer audience-specific words than human content. Moreover, the gap becomes particularly apparent for audiences where humans make greater use of distinctive vocabulary: LLMs often fail to increase their audience-specific vocabulary to a comparable level. These results suggest that, even when explicitly conditioned on a target audience, LLMs only partially reproduce the audience-specific lexical choices observed in human creation.

4.1.3 Human Evaluation

We further conduct a human evaluation to assess whether human readers perceive an audience-design gap between HGC and `gpt-5.4-mini`. Annotators rate each post along five dimensions: tone fit, contextual usefulness, emotional resonance, non-template originality, and engagement potential. The evaluation procedure is provided in Appendix D.5.

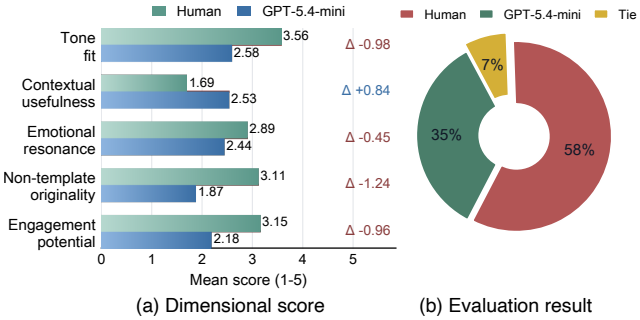


Figure 8: Human evaluation of audience design ability. (a) Average scores on five audience-oriented dimensions. (b) Pairwise results show that HGC is preferred more often than `gpt-5.4-mini` for the same target audience.

As shown in Figure 8 (a), human-written content outperforms `gpt-5.4-mini` on four of the five dimensions, with the largest gaps in *non-template originality* and *engagement potential*. This suggests that human creators produce audience-specific expressions that are less formulaic and more likely to attract interaction.

Human content also shows stronger tone fit and emotional resonance, reflecting better adaptation to the communicative style and affective expectations of the target audience. By contrast, `gpt-5.4-mini` leads only in contextual usefulness: real UGC often serves experiential, emotional, or resonance-building purposes, while LLM-generated posts tend to provide task-oriented steps, summaries, suggestions, and conditional advice such as `if your budget is limited, you can....`

Pairwise comparison in Figure 8(b) further supports this conclusion: HGC is preferred in 58% of matched comparisons, while `gpt-5.4-mini` is preferred in 35%, with 7% ties. Overall, LLM-generated content can be useful and fluent, but remains weaker in distinctive, emotionally resonant, and engagement-oriented audience design.

5 Related Works

Audience-Centric Evaluation. As a central concept in sociolinguistics and pragmatics, the audience has long been recognized as a key factor shaping language use and content adaptation (Bell, 1984; Nguyen et al., 2016). Recent work has explored audience-centric modeling in several generation and adaptation tasks, including text simplification for specific reader groups (Scarton & Specia, 2018), audience-centric style infusion (Moorjani et al., 2022), socially-aware question generation (Stewart & Mihalcea, 2022), and personalized persuasive message generation (Ding & Pan, 2016; Al Khatib et al., 2020).

With the rise of LLMs, audience-centered evaluation has gained increasing attention. Rooein et al. (2023) evaluate LLM adaptation to demographic audiences with different ages and education levels. Wang et al. (2024) constructs a user-centric multi-intent benchmark from real-world LLM use cases, and Liu et al. (2026) studies the benchmark-user need mismatch in climate-change evaluation. Unlike these user-centered benchmarks, our work focuses on the audience-oriented content creation gap between humans and LLMs and its impact on the platform ecosystem.

Impact of Generative AI on Content Ecosystem. As AIGC scales rapidly, it becomes urgent to examine its impact on the platform ecosystem Wei & Tyson (2024). Some studies have explored the influence of AIGC on information retrieval (IR) systems from a model-level perspective Dai et al. (2024). Wang et al. (2025) shows that AIGC with low perplexity can induce source bias in IR models, making retrievers more likely to expose AIGC under system feedback loops Zhou et al. (2025).

Beyond model-level impact, recent studies have used platform-scale data to examine how AIGC affects social media ecosystems, including content creation and consumption patterns in Pixiv (Wei & Tyson, 2024) and temporal AIGC dynamics in RedNote (Li et al., 2026). Our work complements these studies by offering an audience-centered perspective, examining how LLMs' creation strategies may shape content ecosystems. Compared with prior economic simulation studies of human-AI content ecosystems (Yao et al., 2024a; Zhang & Zhang, 2024; Raghavan, 2025; Wang et al., 2026), our simulation is grounded in empirically observed audience-oriented gaps rather than stylized supply or quality primitives. Rather than reproducing real-world platform dynamics, we use simulation as a mechanism-oriented tool to study how audience-strategy gaps may be amplified through repeated platform competition.

6 Conclusion

We study human–LLM gaps in audience-oriented content creation using AudCraft, a large-scale dataset grounded in real-world platform data. For *Audience Orientation*, we find that LLMs concentrate their generations on a narrower range of audience segments than human creators. Our competition-based simulation further suggests that such concentrated orientation may be amplified under competition for limited audience attention, leading to increasingly uneven content supply and intensified competition with human creators. For *Audience Design*, we find that LLMs respond to audience information but do not fully reproduce human-like adaptation, exhibiting weaker and only partially aligned audience-specific patterns. Together, these findings reveal systematic gaps in both whom LLMs create for and how they adapt to different audiences, highlighting the importance of audience-oriented evaluation for understanding the broader impact of LLM-generated content on content ecosystems.

Ethical Considerations

This work builds upon the publicly released *Qilin* dataset (Chen et al., 2025), a multimodal information retrieval dataset collected from Xiaohongshu. We do not crawl or collect additional UGC from the platform. Instead, our dataset is constructed by filtering and reorganizing the existing *Qilin* note corpus for audience-oriented content creation analysis.

Qilin has already applied several privacy and safety protection procedures before release. According to its documentation, the original dataset construction includes filtering unsafe textual and visual content, removing content involving private information of ordinary individuals, and filtering images containing real human faces. *Qilin* also provides anonymized identifiers and encrypted user features rather than directly identifiable user identities. In our work, we further restrict the use of *Qilin* to research purposes and focus on note-level textual content and aggregate audience-category statistics. We do not attempt to identify users, recover individual user trajectories, infer sensitive attributes, or profile specific creators.

Finally, any release of derived resources from this work should respect the license and usage terms of the original *Qilin* dataset. We will not redistribute raw *Qilin* records outside the conditions of the original dataset release. The purpose of this work is not to detect, penalize, or stigmatize human creators or AI-assisted creators, but to better understand how large-scale AI-assisted content creation may affect audience coverage, content diversity, and creator competition.

AI Use Statement

We used generative AI tools in two ways. First, GPT-5.5 was used as part of the research pipeline to assist with taxonomy coverage analysis. These outputs were reviewed and validated by human annotators or the authors according to the procedures reported in the paper. Second, we used generative AI tools, including ChatGPT, for language polishing, sentence rewriting, and clarity improvement.

Generative AI tools were not used to independently generate the paper's research questions, make final methodological decisions, interpret results, or determine the scientific claims. All AI-assisted text and research outputs were reviewed and edited by the authors, who take full responsibility for the final manuscript.

References

Hervé Abdi. Coefficient of variation. *Encyclopedia of research design*, 1(5):169–171, 2010.

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions: H. abdollahpouri et al. *User Modeling and User-Adapted Interaction*, 30(1): 127–158, 2020.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7067–7072, 2020.

Nasser A Alsadhan. Decoding ai authorship: can llms truly mimic human style across literature and politics? *Digital Scholarship in the Humanities*, pp. fqag040, 2026.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, volume 2024, pp. 24814–24836, 2024.

Allan Bell. Language style as audience design. *Language in society*, 13(2):145–204, 1984.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988.

Etienne Brunet. *Le vocabulaire de Jean Giraudoux: Structure et évolution*. Slatkine, 1978.

Jia Chen, Qian Dong, Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, et al. Qilin: A multimodal information retrieval dataset with app-level user sessions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3670–3680, 2025.

Yu-Ta Chen, Yaw-Huei Chen, and Yu-Chih Cheng. Assessing chinese readability using term frequency and lexical chain. *Computational Linguistics and Chinese Language Processing*, 18(2):1–18, 2013.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 526–537, 2024.

Ramtin Davoudi, Kartik Thakkar, Nazanin Donyapour, Tyler Derr, and Hamid Karimi. Assessing capabilities of large language models in social media analytics: A multi-task quest. *arXiv preprint arXiv:2604.18955*, 2026.

Tao Ding and Shimei Pan. Personalized emphasis framing for persuasive message generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1432–1441, 2016.

Michael H. Goldhaber. The attention economy and the net. *First Monday*, 2(4), 1997.

Hen-Hsen Huang and Hsin-Hsi Chen. Interpretation of chinese discourse connectives for explicit discourse relation recognition. In *Proceedings of COLING*, 2014.

Fred Jelinek, Robert L. Mercer, Lalit R. Bahl, and James K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63, 1977.

Evangelos Katsamakas and Juan Manuel Sánchez-Cartas. Generative artificial intelligence, content creation, and platforms. *Journal of Industry, Competition and Trade*, 24, 10 2024. doi: 10.1007/s10842-024-00430-9.

Mike Kestemont. Function words in authorship attribution: From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, pp. 59–66, 2014.

Sueyoul Kim, Ginger Zhe Jin, and Eungik Lee. Does generative ai crowd out human creators? evidence from pixiv. Technical report, National Bureau of Economic Research, 2026.

Kai A. Konrad. *Strategy and Dynamics in Contests*. Oxford University Press, 2009.

Jason MS Lam, Hishamuddin Ismail, and Sean Lee. From desktop to destination: User-generated content platforms, co-created online experiences, destination image and satisfaction. *Journal of Destination Marketing & Management*, 18:100490, 2020.

Jiasun Li. Is generative ai an existential threat to human creatives? insights from financial economics. *arXiv preprint arXiv:2407.19586*, 2024.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pp. 110–119, 2016.

Yudong Li, Yufei Sun, Peiru Yang, Yuhan Yao, Wanyue Li, Jiajun Zou, Haoyang Yang, Haotian Gan, Linlin Shen, and Yongfeng Huang. Rednote-vibe: A dataset for capturing temporal dynamics of ai-generated text in lifestyle social media, 2026. URL https://arxiv.org/abs/2509.22055.

Rui Lin. Quantitative research on readability formula of texts for chinese as a second language. *International Journal of Learning and Teaching*, 9(3), 2022.

Tao Lin, Kun Jin, Andrew Estornell, Xiaoying Zhang, Yiling Chen, and Yang Liu. User-creator feature polarization in recommender systems with dual influence. *Advances in Neural Information Processing Systems*, 37:125663–125701, 2024.

Eden Litt and Eszter Hargittai. The imagined audience on social network sites. *Social Media+ Society*, 2(1):2056305116633482, 2016.

Oucheng Liu, Lexing Xie, and Jing Jiang. Llm benchmark-user need misalignment for climate change, 2026. URL https://arxiv.org/abs/2603.26106.

Denis McQuail. *Audience analysis*. Sage, 1997.

Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. Audience-centric natural language generation via style infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1919–1932, 2022.

Dong Nguyen, A Seza Doğruöz, Carolyn Rosé, and Franciska De Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

Umashanthi Pavalanathan and Jacob Eisenstein. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213, 2015.

Ellie Pavlick and Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74, 2016.

Manish Raghavan. Competition and diversity in generative ai, 2025. URL https://arxiv.org/abs/2412.08610.

Swati Rallapalli, Shannon Gallagher, Ronald Yurko, Tyler Brooks, Chuck Loughin, Michele Sezgin, and Violet Turri. Interpretable stylistic variation in human and llm writing across genres, models, and decoding strategies. *arXiv preprint arXiv:2604.14111*, 2026.

Sudha Rao and Joel Tetreault. Dear sir or madam, may i introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of NAACL-HLT*, 2018.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122, 2025.

Qitian Ren. Advertising and content creation on digital content platforms. *Marketing Science*, 43(4): 734–750, 2024.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. Know your audience: Do llms adapt to different age and education levels? *arXiv preprint arXiv:2312.02065*, 2023.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

Carolina Scarton and Lucia Specia. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 712–718, 2018.

Stergios Skaperdas. Contest success functions. *Economic Theory*, 7(2):283–290, 1996a.

Stergios Skaperdas. Contest success functions. *Economic theory*, 7(2):283–290, 1996b.

Wendell R Smith. Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21(1):3–8, 1956.

Ian Stewart and Rada Mihalcea. How well do you know your audience? toward socially-aware question generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 255–269, 2022.

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

Ankur Tiwari. Generative ai in digital content creation, curation and automation. *International Journal of Research Science and Management*, 10(12):40–53, 2023.

Nafis Irtiza Tripto, Saranya Venkatraman, Mahjabin Nahar, and Dongwon Lee. Beyond checkmate: exploring the creative chokepoints in ai text. *arXiv preprint arXiv:2501.19301*, 2025.

Gordon Tullock. Efficient rent seeking. In James M. Buchanan, Robert D. Tollison, and Gordon Tullock (eds.), *Toward a Theory of the Rent-Seeking Society*, pp. 97–112. Texas A&M University Press, 1980.

Gordon Tullock et al. Efficient rent seeking. *Toward a theory of the rent-seeking society*, 97:112, 1980.

Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32:323–352, 1998.

Haoyu Wang, Sunhao Dai, Haiyuan Zhao, Liang Pang, Xiao Zhang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. Perplexity trap: Plm-based retrievers overrate low perplexity documents. *arXiv preprint arXiv:2503.08684*, 2025.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3588–3612, 2024.

Ziming Wang, Jingyan Li, Wenyi Zhang, and Xiaowei Guo. When capability hurts: Strategic adoption and disclosure of artificial intelligence in creator competition. *Expert Systems with Applications*, 311:131406, 2026. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2026.131406. URL https://www.sciencedirect.com/science/article/pii/S0957417426003192.

Yiluo Wei and Gareth Tyson. Understanding the impact of ai generated content on social media: The pixiv case, 2024. URL https://arxiv.org/abs/2402.18463.

Wei Xiong, Michael Recce, and Brook Wu. Intent-based user segmentation with query enhancement. *International Journal of Information Retrieval Research (IJIRR)*, 3(4):1–17, 2013.

Chen Xu, Xiaopeng Ye, Jun Xu, Xiao Zhang, Weiran Shen, and Ji-Rong Wen. Ltp-mmf: Toward long-term provider max-min fairness under recommendation feedback loops. *ACM Transactions on Information Systems*, 43(1):1–29, 2024.

Nianwen Xue. Annotating discourse connectives in the chinese treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II*, 2005.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-k recommendation under competing content creators? In *International conference on machine learning*, pp. 39674–39701. PMLR, 2023.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Human vs. generative ai in content creation competition: symbiosis or conflict? *arXiv preprint arXiv:2402.15467*, 2024a.

Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Jingzhou Liu, Qifan Wang, Haifeng Xu, and Hongning Wang. User welfare optimization in recommender systems with competing content creators. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3874–3885, 2024b.

Xiaopeng Ye, Chen Xu, Jun Xu, Xuyang Xie, Gang Wang, and Zhenhua Dong. Guaranteeing accuracy and fairness under fluctuating user traffic: A bankruptcy-inspired re-ranking approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2991–3001, 2024.

Yukun Zhang and Tianyang Zhang. The impact of generative ai on content platforms: A two-sided market analysis with multi-dimensional quality heterogeneity. *arXiv preprint arXiv:2410.13101*, 2024.

Zoie Zhao, Sophie Song, Bridget Duah, Jamie Macbeth, Scott Carter, Monica P Van, Nayeli Suseth Bravo, Matthew Klenk, Kate Sick, and Alexandre LS Filipowicz. More human than human: Llm-generated narratives outperform human-llm interleaved narratives. In *Proceedings of the 15th Conference on Creativity and Cognition*, pp. 368–370, 2023.

Yuqi Zhou, Sunhao Dai, Liang Pang, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. Exploring the escalation of source bias in user, data, and recommender system feedback loop. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1676–1686, 2025.

APPENDIX

A LIMITATIONS

This work has several limitations. First, AUDCRAFT is constructed from a Chinese content platform, and our empirical analysis is therefore grounded in Chinese-language social media contexts. Audience orientation may vary across languages, cultures, and platform norms, so future work should extend the dataset and analysis to multilingual and cross-cultural settings.

Second, this paper mainly studies how LLMs differ from human creators in audience strategies for textual content creation. However, recent advances in multimodal generative models have enabled the scalable production of images, videos, audio, and other forms of media content. Audience orientation in multimodal creation may involve different signals and mechanisms, such as visual style, sound, editing, and cross-modal presentation. Future work can extend our framework beyond text to examine how generative models target and adapt to audiences in multimodal content creation.

Third, our simulation abstracts away many complexities of real platform ecosystems. Although it helps reveal potential mechanisms through which LLMs may affect creator competition and audience markets, real platforms involve more complex recommendation algorithms, user feedback loops, creator strategies, and policy interventions. Future work can build more realistic simulation environments grounded in platform-scale behavioral data.

Finally, our study mainly compares human-created and LLM-generated content under controlled generation settings. In practice, many creators may use LLMs as assistive tools rather than fully automated content producers. Understanding hybrid human–AI creation remains an important direction for future research.

B DATASET CONSTRUCTION DETAILS

B.1 DETAILS OF THE MINING OF IN-THE-WILD HUMAN-WRITTEN POSTS

To understand how human creators produce content for diverse audiences in real-world content platforms, we construct AUDCRAFT by carefully curating human-written posts from *Qilin* (Chen et al., 2025), a large-scale collection of real-world user-generated content. Our goal is to obtain a clean corpus that preserves naturally occurring human creation while minimizing low-quality, duplicated, corrupted, and potentially AIGC.

We begin with 1,983,938 raw posts and apply a sequence of lightweight filtering steps. Specifically, we remove posts that (1) have a missing title or body; (2) contain fewer than 30 visible non-whitespace characters; (3) contain no textual content or have a Chinese-character ratio below 50%; (4) are exact duplicates of other posts; or (5) contain more than 5% anomalous characters. These filters are designed to remove samples with insufficient textual information or obvious data-quality issues while otherwise preserving the diversity of naturally occurring content. After these filtering steps, 1,846,979 posts remain.

Since our goal is to use AUDCRAFT as a human-written reference, an additional concern is potential contamination by AIGC. We therefore apply a recent AIGC detector (Bao et al., 2024) to the remaining posts and exclude samples classified as likely AI-generated. The detector identifies 256,730 posts as suspected AIGC, which are removed from the corpus. This leaves 1,590,249 posts in the final human-written reference corpus. Table 2 reports the number of posts removed and retained at each filtering step. The resulting AUDCRAFT corpus serves as our human-written reference for subsequent comparisons with LLM-generated content and for studying how human creators produce content for different audiences.

B.2 DETAILS OF AUDIENCE TAXONOMY CONSTRUCTION

We construct the audience taxonomy through a semi-automatic process that combines human induction with LLM-assisted coverage expansion. Rather than asking an LLM to generate audience categories from scratch, we first derive seed categories from real-world human-created content and

Table 2: Filtering statistics for the human-written reference corpus.

Filtering criterion	Removed	Posts retained
Raw posts	–	1,983,938
Missing title or body	123	1,983,815
Fewer than 30 visible non-whitespace characters	116,865	1,866,950
Chinese-character ratio below 50% or no text	16,383	1,850,567
Exact duplicate text	3,574	1,846,993
More than 5% anomalous characters	14	1,846,979
Suspected AI-generated content	256,730	1,590,249

then use an LLM to identify potentially uncovered audience types. The construction process consists of four stages.

Stage 1: Human induction of seed categories. We first randomly sample 100 HGCs from AUDCRAFT. Ten human annotators inspect each sampled post together with its associated hashtags and engagement-positive search queries, which provide complementary signals about the audiences interested in the content. For each post, annotators characterize both *who* is likely to be interested in the content and *why* they would consume it. They then group recurring audience needs into tentative audience domains and intents. This initial induction produces 17 candidate audience domains and 14 candidate audience intents.

Stage 2: Human refinement of seed categories. The annotators then jointly review and refine the initial categories. Specifically, they *merge* categories that capture substantially overlapping audience needs, *split* categories that are overly broad, *remove* labels that are too topic-specific to represent reusable audience types, and *clarify* ambiguous category definitions and boundaries. We favor categories that recur across different content topics rather than labels tied to particular entities or events. After this refinement, the taxonomy is consolidated into 9 audience domains and 10 audience intents.

Stage 3: LLM-assisted coverage expansion. To examine whether the human-derived taxonomy generalizes beyond the initial 100 posts, we use `gpt-5.5` to analyze an additional 1,000 HGCs sampled from AUDCRAFT. The model is provided with the existing category names and definitions and asked to assign the most appropriate domain and intent to each post. Importantly, it is not asked to construct a new taxonomy. Instead, when the existing taxonomy cannot adequately characterize the audience of a post, the model is instructed to flag the post as potentially uncovered and provide a short description of the missing audience type.

Among the 1,000 posts, 54 (5.4%) are flagged as potentially uncovered. We aggregate these cases and return them to the human annotators for further review. A new category is considered only when multiple flagged examples exhibit a recurring audience need that cannot be reasonably captured by an existing category; isolated or highly topic-specific cases do not trigger taxonomy expansion.

Stage 4: Human review and taxonomy finalization. The human annotators manually review the 54 flagged cases and determine whether they reveal genuinely missing audience types or can be accommodated by refining existing category definitions. Based on this review, the taxonomy is further expanded and consolidated into its final form of **12 audience domains and 12 audience intents**. Thus, `gpt-5.5` serves only as a scaling mechanism for identifying potentially uncovered cases, while all decisions about adding, merging, removing, or redefining categories are made by human annotators.

The resulting taxonomy therefore combines bottom-up induction from real-world human content with broader coverage checking over additional samples. The complete taxonomy, together with category definitions and representative examples, is provided in Table 3.

Table 1 presents representative cases of audience segmentation in AUDCRAFT. Each row shows how a real platform post is associated with an audience domain and an audience intent, together forming an audience segment. For each case, we provide the post title, excerpt, creator hashtags, and engagement-related search queries. The English translations are provided for readability, while the

Table 3: Audience domain and audience intent labels used in AUDCRAFT.

	Audience Domain		Audience Intent
1	Lifestyle Consumers	1	Learning from Experience
2	Tech Users	2	Following Practical Steps
3	Wellness Seekers	3	Understanding Concepts
4	Home & Living Enthusiasts	4	Solving a Problem
5	Food Enthusiasts	5	Making a Purchase Decision
6	Travel Explorers	6	Exploring Viewpoints
7	Rights Seekers	7	Seeking Emotional Connection
8	Career Builders	8	Validating a Choice
9	Knowledge Seekers	9	Keeping Updated
10	Culture Enthusiasts	10	Engaging with a Story
11	Relationship Seekers	11	Reflecting on Personal Meaning
12	Self-Improvers	12	Scanning Key Information

original Chinese text preserves platform-native expressions. These examples illustrate how audience segments are grounded in real content, user needs, and platform engagement signals.

C More Evaluation Experiments

C.1 Audience Orientation under Varying Decoding Temperature

Figure 9 examines whether the observed audience-orientation concentration is sensitive to decoding temperature. We vary the temperature from 0.7 to 1.3 while keeping other decoding parameters fixed. Across all three settings, the generated content remains highly concentrated in a small number of audience domain–intent segments. The dominant segments are stable: *Learning from Experience* and *Seeking Emotional Connection* mainly concentrate on *Food Enthusiasts*, *Travel Explorers*, and *Self-Improvers*, while most other domain–intent combinations receive near-zero mass. For example, under $\tau = 1.0$, the largest cells are still concentrated around *Learning from Experience* $\times$ *Food Enthusiasts/Travel Explorers* and *Seeking Emotional Connection* $\times$ *Self-Improvers*. Increasing the temperature to 1.3 slightly changes local proportions, but does not produce broader audience coverage.

These results suggest that the audience-orientation gap is not simply an artifact of using a low-temperature decoding strategy. Even under a higher-temperature setting, LLM-generated content continues to occupy a narrow subset of audience segments, indicating that homogenized audience imagination is a relatively stable generation tendency.

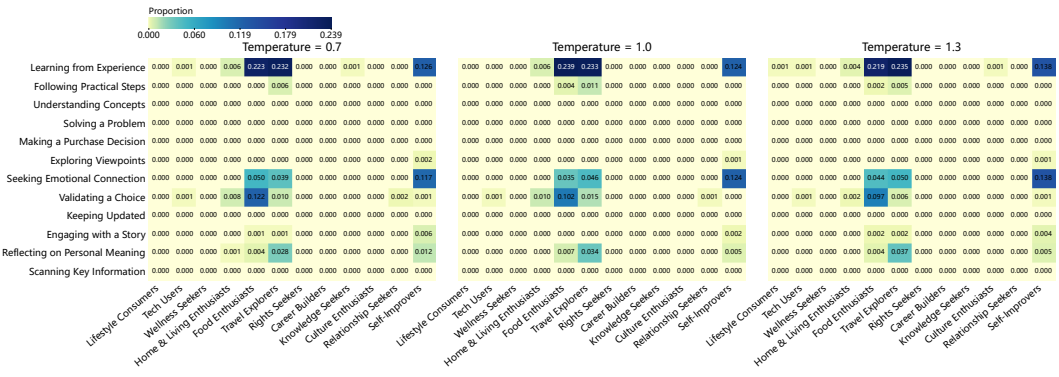


Figure 9: Robustness of audience-orientation concentration under different decoding temperatures with creator persona prompting strategy. Varying temperature from 0.7 to 1.3 does not substantially reduce concentration, suggesting that the observed pattern is not an artifact of a specific decoding setting.

D Evaluation Framework Details

D.1 Experiment Setups

For all HuggingFace models, generations are performed on NVIDIA A6000 GPUs. For closed-source models, we use their respective APIs to obtain responses. We use `gpt-5.5-2026-04-23` [2] solely to assist in expanding the coverage of the human-constructed taxonomy by flagging potentially uncovered audience types, which are subsequently reviewed by human annotators. For evaluation, we use `gpt-5.4-mini-2026-03-17`, `deepseek-ai/DeepSeek-V3.2` [3], `qwen3.5-plus-2026-04-20` [4], `qwen3-max-2026-01-23`, `qwen3.5-35B`, `MiniMax-M2.5` [5], `kimi-k2.5` [6], and `glm-5` [7]. All models are evaluated in chat-completion mode with the same system instruction and prompt templates. Unless otherwise specified, we set the temperature $\tau = 1$, top-$p = 1$, frequency penalty to 0, presence penalty to 0. Audience labels used for the audience-orientation distributions are assigned by human annotators.

D.2 Audience Orientation Experiment Details

D.2.1 Human Annotation Details

Each post is independently annotated by three annotators sampled from a pool of 10 annotators, who assign one audience domain and one audience intent according to the fixed taxonomy. Inter-annotator agreement is high, with Fleiss' $\kappa = 0.91$ for audience domain and 0.85 for audience intent. For each dimension, the majority label among the three annotations is taken as the final label, while cases without a majority agreement are reviewed by an additional annotator for adjudication. Annotators receive only the post title, body, and taxonomy definitions and are blinded to the content source and model identity.

D.2.2 Audience Orientation Prompt

For the audience orientation experiment, we use an open-ended platform-style content creation prompt. The prompt does not specify any audience domain or audience intent.

Minimal Prompt

```
Write a Chinese user-generated content post.

The post should include:
1. a title,
2. a main body,
```

Platform-Style Prompt

```
Please write a Chinese post that could appear on a user-generated content platform.

The post should include:
1. a title,
2. a main body,


Requirements:
- Freely choose the topic, intended audience, and writing style.
- Do not mention that the content is AI-generated.
- Do not include any explanation outside the post.
```

[2] https://platform.openai.com/docs/models
[3] https://api-docs.deepseek.com
[4] https://www.alibabacloud.com/help/en/model-studio/qwen-api-via-dashscope
[5] https://www.minimax.io/models/text
[6] https://platform.kimi.ai/docs/overview
[7] https://docs.bigmodel.cn/cn/guide/models/text/glm-5

Creator Persona Prompt

```
Assume you are an ordinary creator on a Chinese user-generated content platform.
Please write a natural and original Chinese post.

The post should include:
1. a title,
2. a main body.

Requirements:
- Freely choose the topic, intended audience, and writing style.
- Do not mention that the content is AI-generated.
- Do not include any explanation outside the post.
```

Audience-Domain-Specific Prompt

```
Assume you are an ordinary creator on a Chinese user-generated content platform.
Write a natural and original Chinese post for the following audience domain:

\textbf{Audience domain:} \texttt{\{audience\_domain\}}

\textbf{Audience description:} \texttt{\{audience\_domain\_description\}}

The post should include:
1. a title,
2. a main body.

Requirements:
- Freely choose the topic and writing style within the specified audience domain.
- Make the post suitable for the described audience.
- Do not mention that the content is AI-generated.
- Do not include any explanation outside the post.
```

D.3 Audience Design Prompt

For the audience design experiment, we condition generation on audience segments.

Prompt for Audience Design Evaluation

```
Write one Chinese post according to the specified target audience that could appear on a
     user generated content platform.

Target audience:
- Domain: \{domain\}
- Domain definition: \{domain definition\}
- Intent: \{intent\}
- Intent definition: \{intent definition\}

Output format:
Title: ...
Body: ...


Instructions:
- The post should be tailored to the specified domain and intent.
- Use the target audience information as the only audience cue.
- Do not mention AI generation.
- Do not add any explanation outside the post.
```

D.4 Evaluation Metrics Details

We consider four groups of stylometric metrics to characterize different aspects of the generated and HGC: length and readability, style and discourse, lexical diversity, and audience targeting. Below, we provide the definition and computation procedure for each metric.

D.4.1 METRICS FOR AUDIENCE DESIGN GAP

Spearman correlation. We use Spearman correlation (SPR) to measure whether the LLM cross-audience profile follows the same relative pattern as the human profile. Let $r(s_m^H)$ and $r(s_m^L)$ be the ranks of segment m under the human and LLM profiles, respectively. The Spearman correlation is computed as

$$\mathrm{SPR}(\mathbf{s}^H, \mathbf{s}^L) = \mathrm{Pearson}\big(r(\mathbf{s}^H), r(\mathbf{s}^L)\big),$$

where $r(\mathbf{s}^H)$ and $r(\mathbf{s}^L)$ denote the rank-transformed human and LLM cross-audience profiles. A higher SPR indicates that the LLM profile better preserves the human cross-audience ordering.

Mean Absolute Percentage Error. We use Mean Absolute Percentage Error (MAPE) to measure the absolute deviation of the LLM profile from the human reference:

$$\mathrm{MAPE}(\mathbf{s}^H, \mathbf{s}^L) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left| \frac{s_m^L - s_m^H}{s_m^H + \epsilon} \right|.$$

Here, ϵ is a small constant for numerical stability. Lower MAPE indicates smaller deviation from the human cross-audience profile.

Coefficient-of-variation ratio. To measure the relative strength of cross-audience variation, we use the coefficient-of-variation ratio (CV Ratio) (Abdi, 2010). Given a cross-audience profile $\mathbf{s}$, its coefficient of variation is defined as

$$\mathrm{CV}(\mathbf{s}) = \frac{\sigma(\mathbf{s})}{\mu(\mathbf{s}) + \epsilon_1},$$

where $\mu(\mathbf{s})$ and $\sigma(\mathbf{s})$ denote the mean and standard deviation of the profile values across audience segments, and ϵ_1 is a small constant for numerical stability. We then define the CV Ratio between LLM-generated content and HGC as

$$\text{CV-Ratio} = \frac{\mathrm{CV}(\mathbf{s}^L)}{\mathrm{CV}(\mathbf{s}^H) + \epsilon_2}.$$

A CV Ratio below 1 indicates that LLM-generated content varies less than HGC across audience segments, suggesting weaker audience-specific adaptation. In our experiments, we set $\epsilon = \epsilon_1 = \epsilon_2 = 10^{-8}$.

D.4.2 LENGTH AND READABILITY

Text length measures the overall amount of textual content (Tweedie & Baayen, 1998). In our implementation, it is computed as the number of non-whitespace characters in a text. This removes spaces and line breaks while preserving the substantive characters in the text.

Average sentence length measures how much textual content is packed into each sentence (Chen et al., 2013; Lin, 2022). We first split each text into sentences using sentence-ending punctuation. The metric is then computed as the text length divided by the number of detected sentences.

Formality score measures the degree to which a text is written in a formal style (Pavlick & Tetreault, 2016; Rao & Tetreault, 2018). We split each text into sentences and apply a Chinese formality classifier to each sentence. The classifier returns the probability that a sentence is more formal. The document-level formality score is the average of these sentence-level probabilities. Higher values indicate a more formal writing style.

Readability is measured using a Chinese readability formula designed to estimate reading difficulty from sentence-level textual features (Chen et al., 2013; Lin, 2022). For each text, we clean the text, split it into sentences, and pass the resulting sentence list to the readability function. The resulting score summarizes the estimated reading difficulty of the text.

Perplexity is computed from the next-token probabilities estimated by a reference causal language model `qwen3-8B`[8]. Given a tokenized text $x_i = (w_1, \ldots, w_N)$, we compute the average negative

[8] `https://huggingface.co/Qwen/Qwen3-8B`

log-likelihood of next-token prediction and exponentiate it:

$$\mathrm{PPL}(x_i) = \exp\left(-\frac{1}{N-1}\sum_{t=2}^{N}\log p(w_t \mid w_{<t})\right).$$

Lower values indicate that the text is more predictable under the reference model and more consistent with the model's learned distribution. This is often used as a proxy for fluency (Jelinek et al., 1977; Bengio et al., 2003).

D.4.3 Style and Discourse

Discourse marker ratio measures the frequency of explicit connective and organizational expressions (Xue, 2005; Huang & Chen, 2014). We use a high-precision Chinese discourse marker lexicon with 72 entries across eight functional categories: causal, contrastive, additive, summary, example, comparison, sequential, and conditional markers. The entries are deduplicated, manually checked for category consistency, and matched using longest-first non-overlapping string matching to avoid double-counting nested expressions. The metric is the number of matched discourse markers divided by the number of valid tokens after Chinese word segmentation and punctuation removal.

Modifier ratio measures the relative use of descriptive or qualifying expressions (Biber, 1988). We obtain part-of-speech tags using a Chinese spaCy model and define modifiers as tokens tagged as adjectives or adverbs. The metric is computed as the number of adjective and adverb tokens divided by the total number of non-punctuation tokens.

Function word ratio measures the relative use of grammatical function words in a text (Kestemont, 2014; Tausczik & Pennebaker, 2010). We operationalize function words as tokens assigned to the following POS categories: pronouns (`PRON`), adpositions (`ADP`), particles (`PART`), coordinating conjunctions (`CCONJ`), and subordinating conjunctions (`SCONJ`). For a text x_i, the metric is computed as:

$$\mathrm{FWR}(x_i) = \frac{|\{w \in x_i : \mathrm{POS}(w) \in \mathcal{F}\}|}{|\{w \in x_i : \mathrm{POS}(w) \neq \texttt{PUNCT}\}|},$$

where $\mathcal{F} = \{\texttt{PRON}, \texttt{ADP}, \texttt{PART}, \texttt{CCONJ}, \texttt{SCONJ}\}$. This metric captures structural and grammatical style rather than topic-specific lexical choices.

Question sentence ratio measures the proportion of interrogative sentences in a text (Biber, 1988). We first split each text into sentences and identify question sentences using explicit question marks and implicit Chinese interrogative patterns. The metric is computed as the number of question sentences divided by the total number of sentences.

Internet slang ratio measures the relative frequency of internet slang and netspeak expressions. We build the slang lexicon from Wiktionary's "Chinese internet slang" category[9] and the netspeak category of the Simplified Chinese LIWC dictionary (Tausczik & Pennebaker, 2010). Traditional Chinese forms are converted to simplified Chinese while preserving the original forms, resulting in 1,463 unique raw entries. To reduce false positives in Chinese UGC, we remove overly ambiguous entries such as single Chinese characters, single English letters, single digits, punctuation-only items, and very short ASCII strings. The final implementation uses 1,210 longer terms for non-overlapping substring matching and 51 short terms for exact-token matching. The metric is the number of matched slang expressions divided by the number of valid tokens.

D.4.4 Lexical Diversity.

Distinct-1/2 measure lexical variety at the unigram and bigram levels (Li et al., 2016). After Chinese word segmentation, each metric is computed as the number of unique n-gram types divided by the total number of n-gram tokens, with $n = 1$ for Distinct-1 and $n = 2$ for Distinct-2. Higher values indicate greater lexical diversity.

Brunet Index provides a complementary measure of lexical richness based on the relationship between total token count and vocabulary size (Brunet, 1978; Tweedie & Baayen, 1998). Let N_i

[9] `https://www.wiktionary.org/`

be the total number of valid tokens in text x_i, and let V_i be the number of unique token types. We compute

$$\mathrm{Brunet}(x_i) = \frac{V_i}{N_i^{0.165}}.$$

Under this implementation, higher values indicate greater lexical richness after adjusting for text length.

D.4.5 AUDIENCE TARGETING

Audience vocabulary score measures whether a text uses vocabulary associated with its intended audience segment (Bell, 1984; Nguyen et al., 2016). To avoid leaking model-specific lexical patterns into the reference vocabulary, we construct the audience vocabulary using only HGC. For each audience domain–intent combination g, we pool all HGC texts belonging to g into a grouped document and compute TF-IDF scores (Salton & Buckley, 1988) across all audience segments. Let A_g denote the top 30 words with the highest TF-IDF scores for segment g, which serve as the human-reference audience vocabulary for that segment.

For a text x_i assigned to segment g, let T_i denote its token sequence. We compute

$$\mathrm{AudienceVocabScore}(x_i) = \frac{|\{t \in T_i : t \in A_g\}|}{|T_i|}.$$

Higher values indicate stronger alignment with human audience-specific lexical usage. We apply the same human-derived vocabulary A_g to both HGC and LLM-generated texts, ensuring that the metric does not absorb model-specific lexical patterns from LLM outputs. Thus, the score should be interpreted as a relative alignment with human audience vocabulary, rather than an absolute measure of audience quality.

D.5 HUMAN EVALUATION PROTOCOL

We conduct a human evaluation to assess whether a post is suitable for its target audience. For each annotation instance, three annotators are shown a target audience label, a brief description of the audience, and one candidate post. Annotators are asked to rate the post on a 1–5 Likert scale along the following five dimensions, where 1 indicates "strongly disagree" and 5 indicates "strongly agree".

Tone Fit. The post uses a tone and communication style that are appropriate for the target audience. For example, content targeting trip planners should be clear and informative, while lifestyle-sharing content may be more casual and affective.

Contextual Usefulness. The post provides information, experience, or suggestions that are useful in the audience's likely consumption context. Usefulness does not only refer to factual instructions, but also includes whether the content helps the audience understand, decide, reflect, or engage with the topic.

Emotional Resonance. The post is likely to resonate with the target audience emotionally or socially. It may express feelings, shared experiences, frustrations, aspirations, or values that the audience can relate to.

Non-template Originality. The post avoids overly generic, formulaic, or template-like expressions. It should contain concrete details, distinctive perspectives, or natural expressions that make it feel less standardized.

Engagement Potential. The post is likely to attract the target audience's attention or encourage interaction, such as reading, liking, commenting, sharing, or saving. This dimension captures whether the content is appealing to the intended audience in a platform setting.

Pairwise Preference. In addition to dimensional scoring, annotators are also asked to compare a human-written post and an LLM-generated post associated with the same target audience. The source of each post is hidden, and the order of the two posts is randomized. Annotators choose one of three options: the first post is more suitable, the second post is more suitable, or the two posts are tied.

For each valid target audience segment, we randomly sample 5 matched pairs, where each pair contains one HGC post and one LLM-generated post assigned to the same audience segment. This

Algorithm 1: Tullock-Contest-Based Simulation with AI Entry

Require: Creators $\mathcal{N}$, audience segments $\mathcal{M}$, reward pools $\{V_m\}$, rounds T, entry round T_{AI}, parameters η, β, ρ, and creator costs $\{c_n\}$.
Ensure: Trajectories of segment choices, efforts, rewards, and competition statistics.

```
Initialize creator types, segment choices, efforts, and historical rewards.
for t = 1 to T do
    if t = T_AI then
        Convert selected creators into LLM-assisted creators and update their creation prior.
    end if
    for each creator n ∈ N do
        Sample segment m_n^t from a softmax response over historical rewards.
        Update effort e_n^t by utility gradient if staying in the same segment; otherwise initialize effort.
    end for
    for each segment m ∈ M do
        Let N_m^t = {n : m_n^t = m} and X_m^t = Σ_{j∈N_m^t} e_j^t.
        Allocate rewards R_{n,m}^t = V_m e_n^t / X_m^t for all n ∈ N_m^t and update R̄_{n,m}^t.
    end for
    Record segment-level creator count, total effort, and average reward.
end for
return Simulation trajectories before and after AI entry.
```

Table 4: Mapping between the Tullock contest model and our platform content-competition setting.

Tullock Contest Element	**Platform Interpretation**	**Our Framework**
Contest	Competition for limited resources	Competition for user attention/exposure
Prize V_m	Resource pool in a market	Attention/reward pool of audience segment m
Contestants	Agents competing for the prize	Human and LLM-assisted creators
Contest choice	Which contest to enter	Which audience segment to target
Effort e_n^t	Investment made by contestant n	Content creation number of creator n
Total effort X_m^t	Aggregate competition intensity	Total creator effort in audience segment m
Contest success function	Reward allocation rule	Exposure/reward allocated by relative effort
Reward $R_{n,m}^t$	Payoff received by contestant	Exposure or creator-side reward
Cost $C_n(e_n^t)$	Cost of exerting effort	Creation cost of human or LLM-assisted creators

results in 275 pairs across 55 audience segments. To reduce annotation cost while enabling reliability assessment, we adopt a partially overlapping annotation design. Each pair is primarily assigned to one of three annotators, who are shown the target audience label, its description, and the two anonymized posts in randomized order. In addition, a randomly selected 20% subset of pairs is independently annotated by all three annotators for quality control and agreement estimation.

For the main results, we use the primary annotations for non-overlapping pairs and majority-vote labels for the overlapping subset. We report the proportions of pairs where HGC is preferred, LLM-generated content is preferred, or the two are tied. To assess annotation reliability, we compute Fleiss' κ on the overlapping subset over the three-way preference labels (*HGC preferred*, *LLM preferred*, and *tie*). The inter-annotator agreement is $\kappa = 0.68$, indicating substantial agreement.

E More Simulation Details and Results

The Tullock contest model is adapted to our platform content-competition setting. Table 5 reports the main simulation parameters, Algorithm 1 summarizes the overall workflow of our Tullock-based simulation, including segment selection, effort update, reward allocation, and post-entry dynamics.

E.1 AI-Human Competition as Tullock Contest

The Tullock contest is a classical economic model for studying competition over a limited prize (Tullock, 1980; Skaperdas, 1996a; Konrad, 2009). In this model, contestants invest costly effort, and the prize is allocated according to their relative effort among all competitors. Thus, an agent's reward depends not only on its own effort, but also on the total effort invested by others.

We use this model to abstract competition in platform content ecosystems. On content platforms, user attention and platform exposure are limited resources (Goldhaber, 1997; Yao et al., 2023). Creators targeting the same audience segment, therefore, compete for the same attention pool. Accordingly, each audience segment is treated as a contest, where the prize V_m represents the available attention or reward pool. Human and LLM-assisted creators are the contestants; choosing a contest corresponds to choosing an audience segment; and effort e_n^t corresponds to the amount of content produced by creator n at round t. The total effort X_m^t represents the aggregate content supply, or competition intensity, within segment m.

Table 4 summarizes the mapping between the Tullock contest model and our platform content-competition setting. This abstraction enables us to examine how empirically observed differences in audience targeting and production cost between human and LLM-assisted creators may be amplified through repeated competition.

E.2 Simulation Workflow

E.2.1 Creation Strategy Update

We model creators as adaptive agents. At each round, each creator first selects an audience segment and then updates the production effort invested in that segment. This captures two basic decisions in content creation: which audience to target and how much effort to allocate.

Segment Selection. Creator n selects an audience segment according to a softmax response over historical average rewards. The probability of choosing segment m at round $t+1$ is defined as:

$$P_n^{R,t+1}(m) = \frac{\exp\left(\beta \bar{R}_{n,m}^t\right)}{\sum_k \exp\left(\beta \bar{R}_{n,k}^t\right)}, \tag{2}$$

where $\bar{R}_{n,m}^t$ denotes creator n's average realized reward from segment m up to round t. The parameter β controls the exploration–exploitation trade-off. A larger β makes creators more likely to choose previously rewarding segments, while a smaller β allows broader exploration.

After AI enters the platform, the segment-choice distribution of AI-assisted creators is additionally biased toward the empirical LLM audience distribution P^L. Specifically, after computing the reward-based softmax distribution P_n^R, we update the choice distribution of AI-assisted creator n as

$$P_n^{t+1}(m) = (1-\lambda) P_n^{R,t+1}(m) + \lambda P^{L,t+1}(m), \tag{3}$$

where λ controls the strength of the LLM audience-orientation prior.

Reward Allocation. To model competition for limited user attention, we use the Tullock contest success function Skaperdas (1996b) as the platform's reward allocation rule. Each audience segment m is associated with a fixed reward pool V_m, representing the amount of user attention, exposure, or traffic available in that segment. Creators targeting the same segment compete for this pool according to their relative effort.

At round t, let $\mathcal{N}_m^t$ denote the set of creators targeting segment m,. The reward received by creator n is:

$$R_{n,m}^t = V_m \cdot \frac{e_n^t}{\sum_{j \in \mathcal{N}_m^t} e_j^t}, \quad n \in \mathcal{N}_m^t. \tag{4}$$

Creators update their effort by maximizing a utility that balances exposure reward against production cost. The following proposition gives the resulting effort update rule.

Effort Update. After choosing segment m, creator n receives exposure reward from the segment-level reward pool while paying a convex production cost. The utility is defined as:

$$U_{n,m}^t = R_{n,m}^t - c_n (e_n^t)^\rho, \tag{5}$$

where V_m is the reward pool of segment m, e_n^t is creator n's effort, X_m^t is the total effort of all creators targeting segment m, c_n is the creator-specific cost coefficient, and $\rho > 1$ controls the convexity of production cost.

If creator n remains in the same segment across consecutive rounds, we update its effort by gradient ascent on the utility Yao et al. (2024a):

$$\begin{aligned} e_n^{t+1} &= e_n^t + \eta \frac{\partial U_{n,m}^t}{\partial e_n^t} \\ &= e_n^t + \eta \left[V_m \cdot \frac{X_m^t - e_n^t}{(X_m^t)^2} - c_n \rho (e_n^t)^{\rho-1} \right]. \end{aligned} \tag{6}$$

where η is the learning rate. The first term is the marginal gain from increasing effort under the Tullock allocation rule, while the second term is the marginal production cost.

E.3 More Simulation Result

The detailed parameter settings are reported in Table 5.

Table 5: Main simulation parameters.

Param.	Value	Description
N	10,000	Number of creators
M	120	Number of audience segments
T	3,000	Number of rounds
T_{AI}	1,000	AI entry round
AI rate	20%	Fraction of AI-assisted creators
β	0.5	Softmax response parameter
η	1.0	Effort update learning rate
ρ	2.0	Cost convexity
c_{human}	$\lvert\mathcal{N}(1.0, 0.2^2)\rvert$	Initial human cost
c_{AI}	$0.1 c_{\text{human}}$	AI-assisted cost after entry
λ	0.05	Weight of LLM audience prior
V_m	$\lvert\mathcal{N}(\mu_m, \sigma_m^2)\rvert$	Segment reward from AudCraft

E.3.1 Dynamics of Content Number and Efficiency

As LLM intervenes in the content creation process, the cost of production is substantially reduced. We find that, with the introduction of AI and the resulting decrease in creation cost, the total number of content items on the platform increases sharply, as shown in Figure 10 (a). However, user attention remains limited. When finite attention is distributed across a rapidly expanding content pool, the reward received by each content item, such as exposure, inevitably decreases. This leads to a decline in the average reward efficiency per content item, measured for example by exposure volume, as shown in Figure 10 (b).

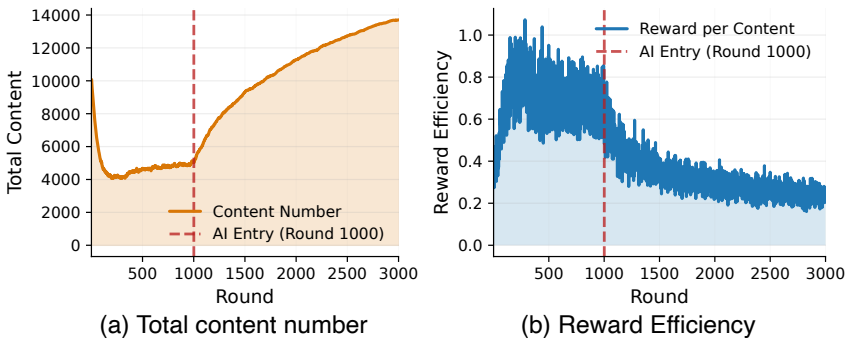


Figure 10: Trend of content number and reward efficiency of content, over simulation rounds.

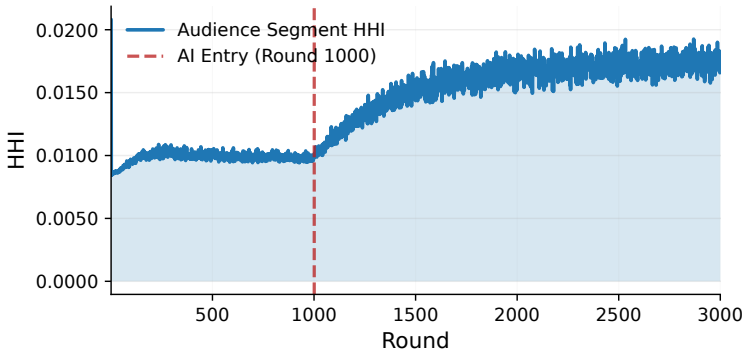


Figure 11: Evolution of audience-segment content concentration, measured by HHI, over simulation rounds.