Leveraging Generative AI for Automated Item Metadata: Enhancing Recommender Performance and Operational Efficiency in a Large-Scale Media Service

1st Hyeong-bin Lee
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
leehb@lguplus.co.kr

1st Yeong-hwan Jeon
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
yeonghwan@lguplus.co.kr

2nd Min-seop Lee
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
oper0815@lguplus.co.kr

3rd Young-suk Moon
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
ysm107@lguplus.co.kr

3rd Hee-cheol Kim
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
hcym0609@lguplus.co.kr

4th Hyun-cheol Jo
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
hcjo@lguplus.co.kr

4th Byoung-Ki Jeon
AX Technology Group
LG Uplus Corp.
Seoul, Republic of Korea
bkjeon@lguplus.co.kr

ABSTRACT

Manual generation of extended item metadata (e.g., mood, emotion) is a costly, inconsistent, and unscalable bottleneck in large-scale media services, hindering recommender quality and operational efficiency. To address this, we present an industrial case study from a large-scale media service 'IPTV' provided by LG Uplus in South Korea, detailing an end-to-end pipeline using generative AI to automatically generate, verify, and evaluate extended metadata. Successfully integrated into the live production recommender system and daily operational workflows, our generative AI-driven approach not only achieved statistically significant improvements in key recommender performance metrics, rigorously validated through large-scale A/B test, but also drastically curtailed the considerable time and resources previously consumed by manual metadata generation efforts. This work demonstrates the practical effectiveness of generative AI for enhancing not only recommender performance but also operational efficiency in a commercial setting.

CCS CONCEPTS

• Information systems → Recommender systems

KEYWORDS

Recommender systems, Large Language Models, Generative AI, Metadata Extraction, Workflow Efficiency

ACM Reference format:

Hyeong-bin Lee, Yeong-hwan Jeon, Min-seop Lee, Young-suk Moon, Hee-cheol Kim, Hyun-cheol Jo, and Byoung-Ki Jeon. 2025. Leveraging Generative AI for Automated Item Metadata: Enhancing Recommender Performance and Operational Efficiency in a Large-Scale Media Service. In *31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), August 3-7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 5 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Item metadata is crucial for enhancing the relevance and diversity of recommendations and mitigating cold-starts [2, 5]. While general metadata (e.g., title, genre) often fails to capture nuanced user preferences, many services utilize extended metadata for subjective attributes (e.g., mood, emotion). This extended metadata offers new discovery opportunities beyond genre or popularity and significantly improves personalization quality.

However, manual extended metadata generation is problematic as it relies on subjective human judgment, leading to significant consistency issues due to discrepancies among taggers [3, 6]. Furthermore, high costs and time create severe scalability constraints, making timely, in-depth tagging infeasible in large-scale service environments like Netflix [1]. These factors persistently hinder recommender system advancement and operational efficiency.

To address this, we developed and deployed an end-to-end automation pipeline using generative AI—a technique recognized for its utility in data annotation and synthesis [7]—in the LG Uplus IPTV recommender system, drawing from the potential of LLMs in recommendations [9]. Our system automatically generates, verifies, and integrates extended metadata. Its verification and evaluation design draws on recent AI content assessment insights [8]. Distinguishing from prior work, our operational end-to-end pipeline features automated generation, multi-stage verification for quality, and seamless system integration. Large-scale A/B test confirmed this approach significantly improves recommender metrics and operational efficiency over manual efforts.

Therefore, this paper presents a practical industrial case study demonstrating simultaneous improvements in both recommender performance and operational efficiency through the effective deployment of generative AI.

Our key contributions include:

(1) An end-to-end automation pipeline and industrial deployment.
(2) Quantified operational efficiency gains.
(3) Validated recommender performance uplift via A/B test.
(4) A practical generative AI application case study.

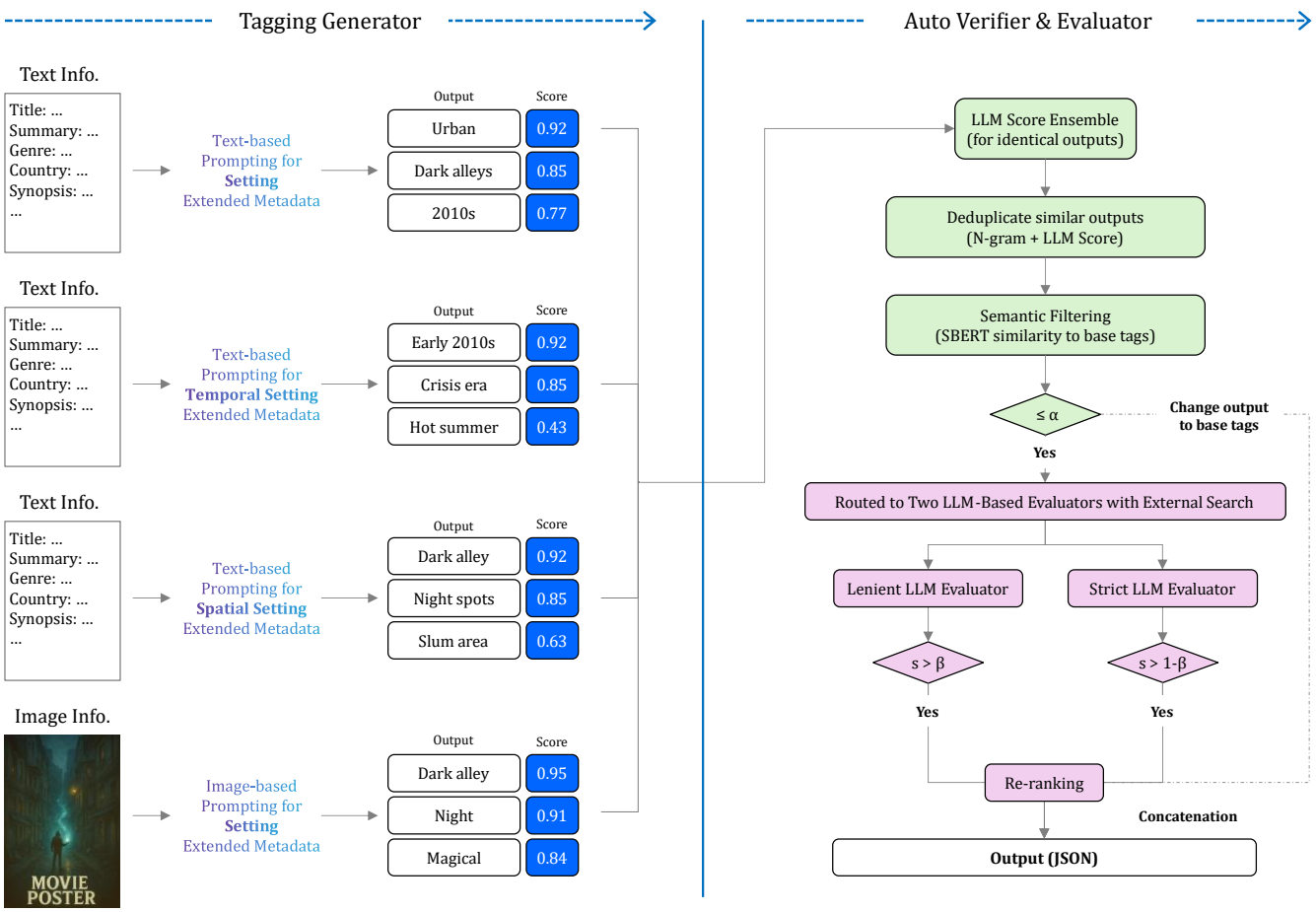

Figure 1: Workflow of the Tagging Generator and Auto Verifier & Evaluator modules. The Tagging Generator uses text/image inputs and prompting to create metadata candidates with scores. Then the Auto Verifier & Evaluator processes these through ensembling, deduplication, semantic filtering, LLM evaluation (Lenient/Strict + External Search), re-ranking, and concatenation to create the final output.

2 Methodology

This section details the overall architecture of the proposed system for automated generation and recommender integration of extended metadata, along with the role and implementation of each component. The entire system is an end-to-end automation pipeline consisting of three main components: the 'Tagging Generator', the 'Auto Verifier & Evaluator', and the 'Recommender System Integrator'. Figure 1 illustrates the detailed workflow of the first two core modules (Generator and Verifier/Evaluator). The 'Recommender System Integrator', described in Section 2.3, integrates the final output into the recommender system.

2.1 Tagging Generator

This module utilizes generative AI to automatically extract various forms of extended metadata candidates [4, 7]. It comprises two parts: one generating extended metadata using only the item's input text, while another part utilizes the input image as well. Prompt engineering techniques, tailored to the characteristics of each extended metadata item, are applied. These include using few-shot examples in the prompts to enhance response consistency and contextual relevance. The generated extended metadata candidates are outputted in a structured JSON format, consisting of its name and reliability score, for subsequent processing and analysis. The generator can be configured based on input data type as follows:

- Text-based Generator: Uses the item's text-based general metadata as input. It combines broad-category prompts capturing the overall characteristics of the item with specific sub-category prompts reflecting subtle nuances or details to generate extended metadata at various levels and perspectives. This strategy is applied across various metadata categories, such as setting and subject matter. This may result in semantically overlapping or similar candidates initially, which are effectively handled in the subsequent verification and evaluation module.
- Image-based Generator: Utilizes the item's poster as input. This is implemented by using a Multimodal LLM that takes both text information and the image as input. This process generates additional extended metadata that is difficult to capture from text alone, such as the item's setting, mood, emotion, and key objects, or core materials visually conveyed by the poster image.

2.2 Auto Verifier & Evaluator

This module refines, verifies, and evaluates the numerous extended metadata candidates received from the Tagging Generator to produce a final output of reliable metadata. The flowchart on the right side of Figure 1 details this process:

• Score Ensemble: If identical metadata candidates are generated through different paths, the scores obtained from each path are ensembled into a single score.

• Deduplicate similar outputs: Candidates that are semantically similar but differently expressed are filtered based on criteria combining N-gram similarity and the ensembled score, keeping only the high-score candidates to eliminate redundancy.

• Semantic Filtering: The remaining candidates are compared against the base tags defined within the established metadata schema (previously defined by humans) using Sentence-BERT-based semantic similarity.

• Schema Alignment Decision: If the similarity score meets or falls below the threshold ($s \leq \alpha$) ('Yes'), then it is considered a novel metadata candidate and proceeds to the next evaluation step. But if it exceeds a specific threshold ($s > \alpha$) ('No'), the candidate is mapped to the most similar base tag, aligning it with the existing schema.

• LLM-based Evaluator Routing: Novel candidates not mapped to the schema are routed to two types of LLM-based evaluators combined with an External Search function for a reliability. Named 'Lenient' and 'Strict', the evaluators reflect distinct thresholds and selection criteria, employ threshold strategies that may seem counter-intuitive at first glance: The Lenient LLM Evaluator judges a candidate as 'Yes' if it passes a relatively higher threshold ($s > \beta$) (as it rigorously assesses candidates to ensure baseline quality for its broader acceptance pool). The Strict LLM Evaluator requires the candidate to pass a lower threshold ($s > 1 - \beta$) (as it applies a more initially permissive bar to consider a wider set of items for its focused scrutiny) to be judged as 'Yes'.

• Re-ranking: The metadata candidates are re-ranked based on a comprehensive consideration of the results from the two evaluators the schema mapping status, and potentially other factors.

• Concatenation: The re-ranked results for novel metadata candidates are combined with the results that were mapped to base tags.

• Final Output: From the final re-ranked metadata list, the Top-K metadata items are output in JSON format to be passed to the next integration stage.

2.3 Recommender System Integrator

The Top-K extended metadata finalized through the Auto Verifier & Evaluator are stored in a central Feature Store along with the item ID, typically in a table format, allowing efficient access for recommender models. This Feature Store houses not only the existing item general metadata but also the high-quality extended metadata generated and verified through this pipeline, kept up-to-date. In the LG Uplus IPTV recommender system, various algorithms (e.g., item-based collaborative filtering, graph-based models) access this Feature Store in real-time or batch mode during prediction or serving to retrieve the necessary item metadata as input features. The entire metadata generation and Feature Store update process runs automatically via batch inference at scheduled intervals and is seamlessly integrated into the existing recommender system infrastructure and data pipelines with minimal changes, enhancing operational efficiency.

3 Experiment

This section presents the evaluation results verifying the effectiveness of the proposed generative AI-based system for automated extended metadata generation and integration, covering offline evaluation, online A/B test for recommender performance/user experience, and operational efficiency improvements. All experiments were conducted based on real data and the live service environment of LG Uplus IPTV.

3.1 Offline Experiment

For evaluation, the extended metadata previously manually tagged for each item was considered the ground truth. For this evaluation, 21,067 items with at least 20 manually tagged extended metadata labels were used as the final offline evaluation dataset. Evaluation metrics used were Precision@K, Recall@K, and MRR@K, with K set to 5, 10, and 25, respectively, calculating average values per item. Additionally, we measured Novelty@K, defined as the average proportion of tags within the Top-K generated list that are not part of the established metadata schema, to evaluate the generation of new tags.

Table 1: Offline Experiment Results. This table shows the results of calculating indicators by considering the existing manual tagging results as the ground truth.

At K	Precision	Recall	MRR	Novelty
@5	0.635	0.394	0.462	0.360
@10	0.594	0.480	0.421	0.417
@25	0.468	0.683	0.404	0.496

The evaluation results showed significant concordance with the existing manual tagging results, achieving an average Precision@25 of 0.468 and Recall@25 of 0.683. This alignment is partly attributed to the 'Schema Alignment' step (Section 2.2), which maps generated candidates to existing tags. Furthermore, the system demonstrated a strong effectiveness in generating novel metadata, as evidenced by the Novelty@K scores. A Novelty@25 score of 0.496 indicates that nearly half of the top 25 generated tags were outside the established schema. However, we recognize that high novelty scores alone do not guarantee the practical relevance or quality of these newly generated tags for enhancing user experience. Therefore, the definitive validation of the effectiveness of the generated metadata set—encompassing both schema-aligned and novel tags—was determined through rigorous online A/B test with real user interactions, as detailed in the next section.

3.2 Online Experiment

To verify the impact on actual recommender performance and user experience, we conducted a 4-week A/B test involving 1,000,000 users in a specific recommender slot (Group A: manual meta, Group B: auto-generated meta). The evaluation metrics included CTR, CVR, PV, UV, and Coverage. Here, Coverage refers to the total number of unique items displayed to users within the specific recommendation slot during the test period. All metric gains are statistically significant at $p < 0.05$.

Table 2: Online Experiment Results. This table shows the A/B test results of applying the existing manual tagging result and the generative AI based tagging result to an algorithm of the LG Uplus IPTV recommender system.

Group	CTR	CVR	PV	UV	Coverage
Group A	X1	X1	X1	X1	X1
Group B	X1.05	X1.09	X1.24	X1.13	X1.28
Ratio	▲5%	▲9%	▲24%	▲13%	▲28%

The A/B test results showed that Group B achieved a 5% increase in Click-Through Rate (CTR) and a 9% increase in Conversion Rate (CVR) compared to the control group [Group A]. Page Views (PV) and Unique Visitors (UV), reflecting actual viewership, increased by 24% and 13%, respectively. This suggests that the auto-generated metadata led to more relevant and engaging recommendations, increasing user interaction and final conversion actions (viewership). Furthermore, the Coverage metric improved by 28%, consistent with the high proportion of novel tags generated by the system identified in the offline analysis (Section 3.1), indicating that users were exposed to a more novel and diverse set of items. For cold-start items (a known challenge [2, 5]), the CTR lift for Group B was even higher at +10.14% compared to the overall average, strongly supporting the effectiveness of the generated extended metadata in enhancing the initial exposure and discoverability of new items with limited user interaction history.

3.3 Operational Efficiency Evaluation

In addition to the offline and online evaluations, improvements in terms of the operational efficiency of the automated system were also analyzed. The time required for generating and verifying extended metadata per item was reduced by 87.3% compared to the previous manual process, and the time needed to apply extended metadata for cold-start items was reduced by 94.5%. This signifies a streamlining of the entire process from an operational perspective, indicating positive impacts on pipeline efficiency across the overall workflow surrounding the recommender system, beyond just the improvements in algorithmic performance.

4 Conclusion and Future Work

This research presented a key industrial case study where an end-to-end automation pipeline, utilizing generative AI to automatically generate, verify, evaluate, and integrate extended metadata into the recommender system, was successfully built and applied to a large-scale commercial recommender platform. The proposed system replaces the inefficient manual metadata generation process, thereby drastically improving operational efficiency while also demonstrating substantial contributions to enhancing recommender quality and user experience. Through this, the study holds significance in successfully validating the practical industrial value of generative AI technology across the two important axes of recommender performance and operational efficiency simultaneously, suggesting the potential for new approaches towards streamlining and advancing overall recommender system and infrastructure operations, aligning with LLM efforts in data annotation [7], label generation [4], and recommendation approaches [9].

Based on the findings of this research, future research and development directions to further advance the system are as follows. First, as the current system primarily focuses on generating extended metadata based on 'static' information available at item creation time, it needs to be expanded to also cover 'dynamic' extended metadata, such as award information or box office performance generated after item release, by actively utilizing external search capabilities within the Tagging Generator stage. Second, to reduce operational time/cost and improve inference speed, applying knowledge distillation techniques—either by replacing the current large-scale generative AI models with lightweight SLM (Small Language Models) or by using the generative AI's outputs as training data to build smaller models specialized for specific tagging tasks—should be considered. Third, to identify extended metadata that contributes to long-term user satisfaction and service metric improvement beyond the effects verified in short-term A/B test, and to refine the system accordingly, research into designing and integrating a Reinforcement Learning-based extended metadata refinement and self-supervised learning loop utilizing user feedback (clicks, filters, favorites, etc.) is important.

BIOGRAPHIES

Hyeong-bin Lee is a Data Scientist with the Personal Agent Tech team in the AX Technology Group at LG Uplus Corp., focusing on recommender systems for media services.

Yeong-hwan Jeon is Team Leader of the Personal Agent Tech team and a Data Scientist in the AX Technology Group at LG Uplus Corp., leading projects related to personalization and recommender systems.

Min-seop Lee is Tech Lead for the Personal Agent Tech team and a Data Scientist in the AX Technology Group at LG Uplus Corp., specializing in the architecture and implementation of recommender systems.

Young-suk Moon is a Data Scientist with the Personal Agent Tech team in the AX Technology Group at LG Uplus Corp., focusing on recommender systems for media services.

Hee-cheol Kim is a Data Analyst with the Personal Agent Tech team in the AX Technology Group at LG Uplus Corp. He focuses on data analysis to support the development and evaluation of recommender systems.

Hyun-cheol Jo is Vice President within the AX Technology Group at LG Uplus Corp., overseeing the development and application of AI technologies for various services.
Byoung-Ki Jeon is Senior Vice President and Head of the AX Technology Group at LG Uplus Corp., leading the group's overall strategy and research in artificial intelligence and data technology.

ACKNOWLEDGMENTS

The movie poster image depicted in Figure 1, used for illustrative purposes in this paper, was generated via a generative AI tool (ChatGPT 4.0). This approach was chosen to demonstrate the concept of image-based input for our metadata generation pipeline while avoiding potential copyright issues associated with the use of actual movie posters. The operational pipeline itself, as described in the paper, is designed to process real-world image information, such as authentic movie posters or trailer thumbnails.

REFERENCES

[1] Xavier Amatriain and Justin Basilico. 2015. Recommender systems in industry: A netflix case study. In Recommender systems handbook. Springer, 385–419.
[2] Yunze Luo, Yuezihan Jiang, Yinjie Jiang, Gaode Chen, Jingchi Wang, Kaigui Bian, Peiyi Li, and Qi Zhang. 2024. Online Item Cold-Start Recommendation with Popularity-Aware Meta-Learning. arXiv preprint arXiv:2411.11225 (2024).
[3] Rock Yuren Pang, Jack Cenatempo, Franklyn Graham, Bridgette Kuehn, Maddy Whisenant, Portia Botchway, Katie Stone Perez, and Allison Koenecke. 2023. Auditing cross-cultural consistency of human-annotated labels for recommendation systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 1531–1552.
[4] Yingchi Pei, Yi Wei Pang, Warren Cai, Nilanjan Sengupta, and Dheeraj Toshniwal. 2024. Leveraging LLM generated labels to reduce bad matches in job recommendations. In Proceedings of the 18th ACM Conference on Recommender Systems. 796–799.
[5] Martin Saveski and Amin Mantrach. 2014. Item cold-start recommendations: learning local collective embeddings. In Proceedings of the 8th ACM Conference on Recommender systems. 89–96.
[6] Andra-Georgiana Sav, Andrew M Demetriou, and Cynthia CS Liem. 2023. Annotation Practices in Societally Impactful Machine Learning Applications: What are Popular Recommender Systems Models Actually Trained On?. In Perspectives@ RecSys.
[7] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. arXiv preprint arXiv:2402.13446 (2024).
[8] Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. arXiv preprint arXiv:2305.13112 (2023).
[9] Junjie Zhang, Ruobing Xie, Yupeng Hou, Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation approach. ACM Transactions on Information Systems (2023).