Enhancing Product Recommendations with Multi-Modal LLMs

Babaniyi Olaniyi – Barcelona, Spain
ISIR-eCom 2025 Workshop (Held with ICDM) – Nov 12, 2025

GitHub: github.com/babaniyi/multimodal_llm_recsys_amazon
Email: horlaneyee@gmail.com

Motivation

- E-commerce data includes text, images, and user interactions.
- Traditional recommenders rely on sparse signals (IDs, ratings).
- Multimodal fusion + LLMs can improve personalization and scalability.
- **Objective:** adapt GPT-2 to predict next items using text and image signals.

Research Question

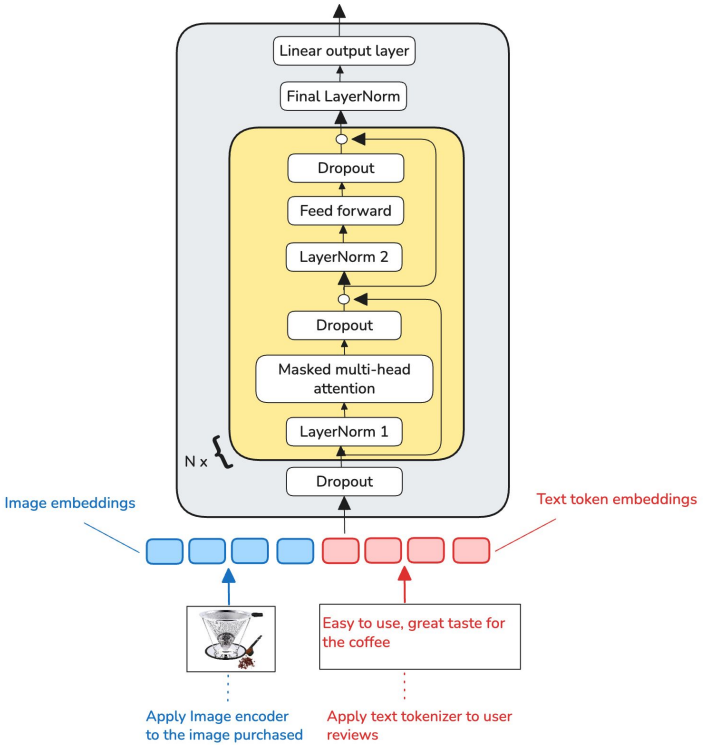
How can we repurpose a large language model (LLM) to:

- Predict the next item in a user’s purchase sequence
- Fuse text and image information efficiently
- Improve accuracy, coverage, and diversity under realistic, resource-constrained training conditions

Model Overview

MM-GPT2Rec = Multimodal Embedding + GPT-2 Decoder

- Text embedding (BPE 768-dim) + Image embedding (ResNet18 256-dim).
- Concatenated and projected to 1024-dim to match GPT-2 hidden size.
- Autoregressive modeling: next product ≈ next word prediction.
- Model learns from sequences of user purchases and predicts the next likely product.



Dataset and Preprocessing

- **Amazon Product Dataset**

  (Appliances, Digital Music, Gift Cards, Health & Personal Care).
- ~400K user-item interactions; users with ≥5 interactions.
- **Text:** review titles + content;

  Image: main product image.
- **Train/Val/Test** = 75/15/10 temporal split; implicit feedback (purchase/review = interaction)

Architecture

- Base: GPT-2 Medium (345M params, 24 layers, 16 heads).
- Input embeddings replaced with multimodal item embeddings (text + image).
- LoRA fine-tuning with tied embeddings; cross-entropy next-item loss.
- Hardware: NVIDIA L4 GPU (24GB VRAM), 8 hours training time. (Ran using Lightning AI platform)

Results Summary

- MM-GPT2Rec: HR@5 = 0.833, Precision@5 = 0.297, Coverage = 0.715, Diversity = 0.749.
- Multimodal baselines (VBPR, DeepCoNN, SASRec): HR@5 ≈ 0.456.
- Traditional methods (Content-Based, MF, CF): HR@5 < 0.1.
- LLM-based model outperforms baselines by 83% in hit rate and doubles catalog coverage.

BASELINE COMPARISON RESULTS AT CUTOFF $k = 5$. HIGHER IS BETTER FOR ALL METRICS.

Model	HR@5	P@5	NDCG@5	MRR	MAP@5	Coverage	Diversity	Novelty
MM-GPT2Rec (Ours)	**0.833**	0.297	0.220	0.207	0.113	**0.715**	0.749	-11.042
Multimodal Baselines								
VBPR	0.456	**0.455**	**0.456**	**0.455**	**0.457**	0.314	0.793	-8.072
DeepCoNN	0.456	**0.455**	**0.456**	**0.455**	**0.457**	0.367	0.793	-8.088
NRMF	0.455	**0.455**	**0.456**	**0.455**	**0.457**	0.310	0.793	-8.086
SASRec	0.456	**0.455**	**0.456**	**0.455**	**0.457**	0.405	0.793	-8.066
Traditional Baselines								
Content-Based	0.080	0.066	0.129	0.075	0.033	0.267	**1.000**	**-3.275**
Matrix Factorization	0.093	0.020	0.039	0.064	0.010	0.056	0.992	-4.677
Hybrid	0.080	0.020	0.051	0.073	0.010	0.242	0.953	-3.528
Collaborative Filtering	0.011	0.002	0.007	0.007	0.001	0.008	0.907	-3.756
Popularity	0.034	0.007	0.006	0.010	0.002	0.000	0.000	-6.237

TABLE II

ABLATION STUDY RESULTS. VALUES REPORTED FROM ACTUAL EXPERIMENTAL RUNS ON FULL DATASI

Configuration	HR@5	Precision@5	MRR	NDCG@5
Text-only	**0.137**	**0.031**	**0.121**	**0.132**
Multimodal (Concatenation)	0.049	0.010	0.046	0.047
Multimodal (Weighted)	0.049	0.010	0.046	0.047
Multimodal (Attention)	0.049	0.010	0.046	0.047

Ablation Insights

Variant	HR@5	MRR
Text-only	**0.137**	0.121
Multimodal (concat / weighted / attention)	0.049	0.046

- Text drives most signal quality; images add noise under naive fusion.
- Image features may add noise; need better cross-modal alignment
- Yet full MM-GPT2Rec outperforms all baselines → transformer learns richer patterns

Qualitative Examples

User purchased

Range Kleen Oven Knob Kit

Next purchases

1. Lifetime Appliance 3362624 Timer Knob Compatible with Whirlpool Washer,
2. WaterSentinel WSG1 Refrigerator Replacement Filter,
3. WaterSentinel WSG-1 Refrigerator Replacement Filter (3-Pack)

Predicted next purchase

1. Dryer Igniter

Prediction is in the same *“appliance repair categories”*

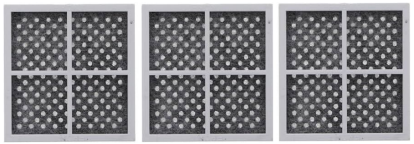
Qualitative Examples

User purchased

1. Mist LT800P LG Refrigerator Water Filter (3 Pack)

Next purchases

1. Tier1 ADQ36006101 Refrigerator Water & Air Filter Combo 3-pk
2. Replacement for LG LT700P Water Filter,
3. NISPIRA Refrigerator Air Filter Compatible with LG LT120F



Predicted next purchases

1. GENUINE Frigidaire 242252702 Valve,
2. GE MWF Refrigerator Water Filter

“Model shows recognition of water filtration system components.”

Traditional models struggle to capture these nuanced relationships as effectively.

MM-GPT2Rec shows category-level and functional understanding of items.

Discussion

- **Strengths:** high hit rate and coverage, cold-start robustness, multimodal reasoning.
- **Weaknesses:** moderate ranking precision, limited novelty, naive image fusion.
- Demonstrates feasibility of LLM-based multimodal recommenders under resource limits.

Limitations & Future Work

- Improve image–text fusion (cross-attention, CLIP-based alignment)
- Scale to larger LLMs (GPT-4, LLaMA-3)
- Explore explainable or conversational recommendation
- Evaluate on other domains (movies, news, health)

Conclusion

- Treating product sequences as language enables contextual, scalable recommendation.
- MM-GPT2Rec captures purchase intent better than traditional models.
- LLMs can power adaptive, context-aware recommendation when paired with multimodal inputs.

→ **MM-GPT2Rec = first step toward foundation-model-based recommendation.**

Thank you

GitHub: github.com/babaniyi/multimodal_llm_recsys_amazon
LinkedIn: @babaniyi