AI and LLM Application at Mercari

Apr 24, 2025

Kaiyi Liu, Search & Discovery Team
kaiyiliu@mercari.com
https://www.linkedin.com/in/kaiyiliu/

AI Engineering Team
Check out our blog: https://ai.mercari.com/en/articles/
and https://engineering.mercari.com/en/blog/

mercari

Agenda

01 About Mercari

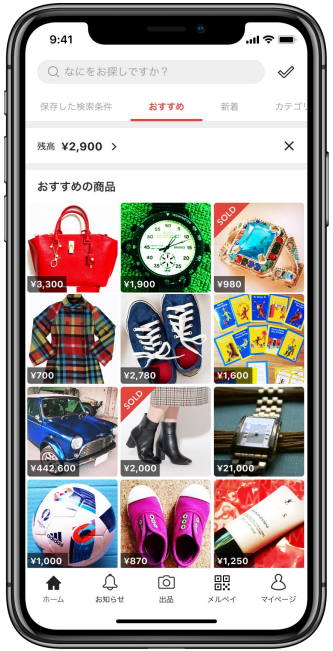
02 We are Open-source Powered Search

03 Core System: Search Workflow and Personalization Features

04 LLM Application in Action

05 Q&A

About Mercari

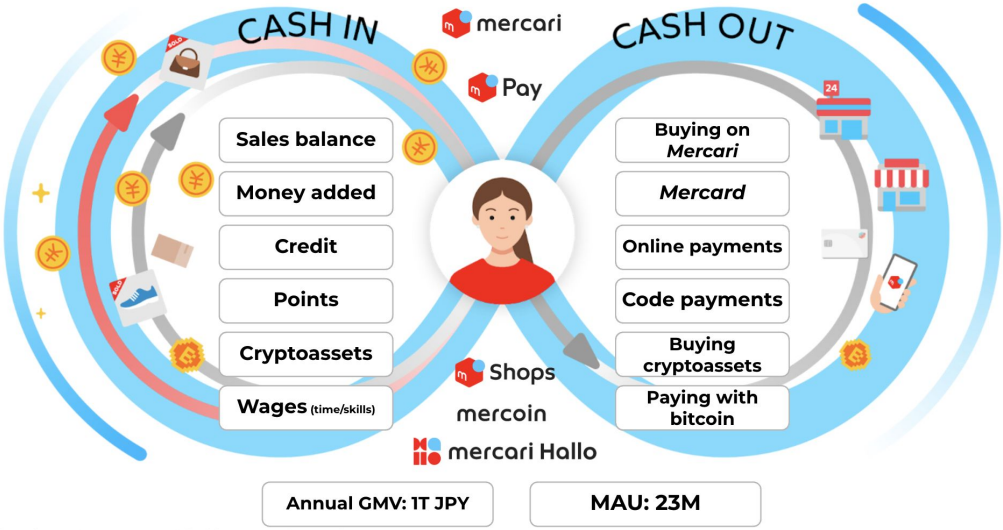


The Mercari app is a C2C marketplace where individuals can easily sell used items.

mercari

About Mercari

- Japan's largest consumer-to-consumer (C2C) online Marketplace
- Circulate **all forms of value** to unleash the **potential** in all **people**



11TH ANNIVERSARY INFOGRAPHICS

Items on *Mercari* Popular Overseas

1 China

1	Games, Toys & Merchandise **Pins & Badges**
2	Games, Toys & Merchandise **Acrylic Standees**
3	Games, Toys & Merchandise **Plushies**
4	Televisions, Audio Devices & Cameras **Digital Cameras**
5	Games, Toys & Merchandise **Comics & Animation**

2 USA

1	Games, Toys & Merchandise **Pokémon Trading Card Game**
2	Fashion **Bags**
3	Games, Toys & Merchandise **Comics & Animation**
4	CDs, DVDs & Blu-rays **K-Pop & Asian Music**
5	Fashion **Tops**

3 Taiwan

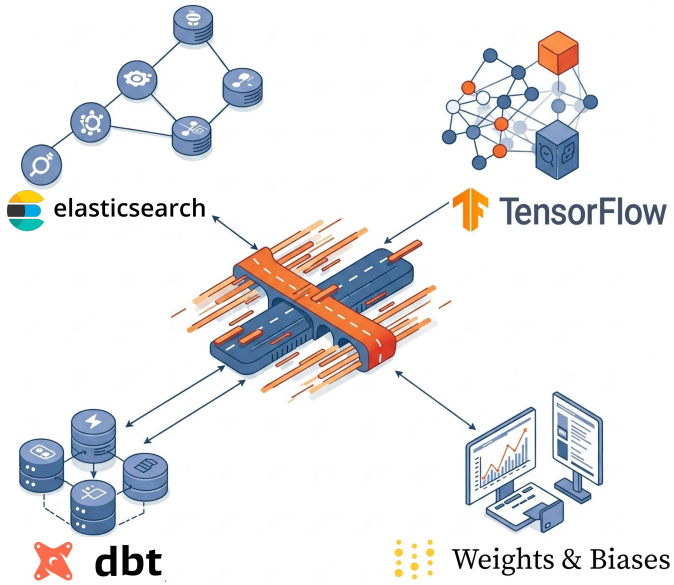
1	Games, Toys & Merchandise **Comics & Animation**
2	Games, Toys & Merchandise **Pins & Badges**
3	Fashion **Bags**
4	Games, Toys & Merchandise **Plushies**
5	Games, Toys & Merchandise **Pokémon Trading Card Game**

Ranking of categories on *Mercari* with high transaction volume from overseas. Data collected between April 1, 2023, and March 31, 2024.

mercari

Source: https://about.mercari.com/en/press/news/articles/20240627_11th_infographics/

Open-source Powered Search



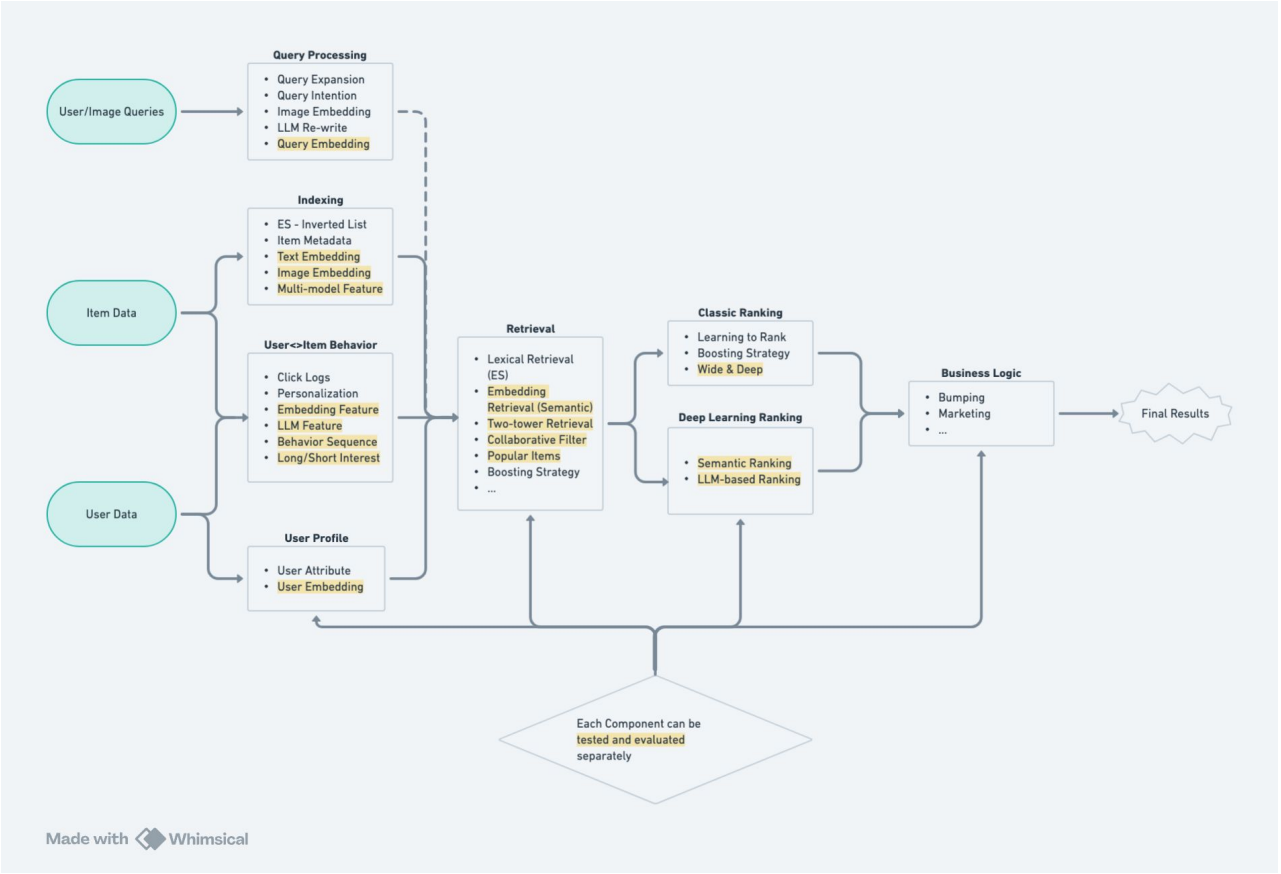
Mercari Ranking System

- Mercari utilizes open-source tools deployed on **GCP** for search ranking.
- **dbt** for data transformation.
- **Elasticsearch** for fast and scalable search for item retrieval.
- **TensorFlow Ranking (**migrating to PyTorch**)** for machine learning models.
- **Weights & Biases** for experiment tracking.
- **Streamlit** for search evaluation UI

Core System: Search Workflow

Future Components

- Offline Evaluation
- Embedding Features



Core System: Personalization Features

In Mercari, we are more focused on leveraging users implicit feedbacks such as likes, views, and purchases to generate user preferences.

01 **Device type targeting (Android, iPhone)**

02 **Ranking based on interaction (view, like, comment)**

03 **Last viewed items for related suggestions (recency)**

04 **User category preferences**

LLM Application in Action: Interview Feedback

- Purpose: Save Interviewers' time in taking & putting interview feedback (we have feedback template to fill in such as language, tech skills and values), and improve the quality of feedback

- This system analyzes interview transcripts using the Gemini API to generate **interview summary** and **identify key points for each of our values**.

- It helps you quickly recall important moments, topics discussed, and candidate responses.

- It intends **NOT** to provide any evaluation or judgement that might create a bias in the interviewers

LLM Application in Action: Synonym Generation

```
Edge cases: note that we treat the following cases as invalid rewrite:
    1. Near synonym: we should NOT expand candidates to phrases that have similar but not identical meaning. (e.g. キャップ ->ハット, カラビナ -> クリップ)
        2. Relaxation: we should NOT mapping to broader meaning (e.g. キャップ -> 帽子, nike air force -> nike, トワツガイ -> 貝, 月姫想本 -> 月姫)
        3. Expanding compound phrases: Phrases like 刺繍ブローチ (刺繍, ブローチ), リバーシブルトートバッグ (リバーシブル, トートバッグ), 映画単行本 (映画, 単行本) should NOT be expanded. Ple
        4. Translate full named entities for works: Phrases like だいじ だいじ どーこだ?, μ"sタペストリー, 金田一少年の事件簿R全巻, 高嶺の花は、散らされたい 上•下 should NOT be translated
    5. Translate a normal phrases to English: Translate phrases like マウンテンジャケット -> mountain jacket should be invalid
        6. Phrases with year, month, day or quantifiers: Phrases like 4体セット, 3個セット should NOT be expanded
    7. Expand short Japanese phrase to a similiar phrase: Expansions like 実力がつく -> 能力向上, カメラ写りません -> カメラが機能しない, 菓子作り -> スイーツ作り should be invalid
    9. Add/Remove stop words to a phrase: Expansions like 開封済アルバム -> 開封済みのアルバム should be invalid
    10. Partial expansions: Partial rewrite of phrases or sentences change the meaning or context. They should be considered as invalid rewrite. e.g. ミニ トート バッグ -> 小
    11. Expand a subset of a full named entity: expanding a subset of a full named entity like "V問題集" in book title "国際関係 講義ノート V問題集" should be invalid

Please carefully check edge cases when generating, as accuracy is much more important to us than coverage.
Your output format is only "[{"source": identified phrase, "target": a list of synonyms, "edge check": check if this expansion meets any edge case rule 1-10}]" form, no

Example:

Text input: always yours Carat盤CDトレカ+weverse特典スングァン
Answer:
[
    {"source": "always", "target": ["オールウェイズ"], "edge check": "meet edge rule 5, invalid"},
    {"source": "Carat", "target": ["カラット"], "edge check": "valid"},
    {"source": "盤CDトレカ", "target": ["盤CD trading card"], "edge check": "meet edge rule 3, invalid"},
    {"source": "トレカ", "target": ["trading card"], "edge check": "meet edge rule 5, invalid"},
    {"source": "weverse", "target": ["ウィバース"], "edge check": "valid"},
    {"source": "スングァン", "target": ["Seungkwan"], "edge check": "valid"}
]
```

synonym

mercari

LLM Application in Action: Synonym Generation

Result: can generate much more synonyms and we are able to connect similar items in our product pools.

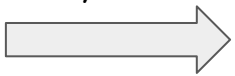
Result in BCR via search +0.61%. Monthly GMV + ~¥105,459,101 for iOS

```
df_old[df_old['source']=="simfree"]
```

✓ 0.0s

	source	target	label
491094	simfree	sim free	0
491095	simfree	sim フリー	0

Pair count increased obviously



```
graph_df[graph_df['source'] == 'simfree
```

✓ 0.0s

	source	target	count
5538	simfree	sim フリー	321
5538	simfree	sim free	321
5538	simfree	sim unlocked	321
5538	simfree	sim ロック 解除	321
5538	simfree	unlocked	321
5538	simfree	sim 無し	321
5538	simfree	sim unlock	321
5538	simfree	sim lock free	321
5538	simfree	sim 未 锁 定	321
5538	simfree	sim 不要	321
5538	simfree	無 sim	321
5538	simfree	シム フリー	321
5538	simfree	sim 未 使用	321
5538	simfree	sim 無 制限	321
5538	simfree	sim なし	321
5538	simfree	sim 無料	321
5538	simfree	sim 自由	321
5538	simfree	sim ロック フリー	321
5538	simfree	sim 无 锁	321

mercari

LLM Application in Action: Query Translation

Flow

- TW user search keyword, SERP with query
- translate to Japanese and search on JP data (by chatgpt 4.0 mini)
- return SERP result
- visit Item page, check if DB has translation
    - If not, Translated and show in Taiwanese (by Gemini 1.5 flash)
        - Store translation into DB
    - If yes and the item listing updated more than 100 characters (TBC)
        - Go back to re-translation and store to DB
        - If yes but not update much: just use it

```
{"role":"system","content":"
* Original text might be a mix of Japanese, Traditional Chinese and English
* Original text is a search keyword for e-commerce marketplace in Japan
* Original text might mean a product series, model name, brand name, idol and character name
* Your task is to translate it to the correct Japanese
* If you cannot translate it reply just N/A
* If translation is not needed just reply the original text
"},
{"role":"user","content": "單速車"}]
```

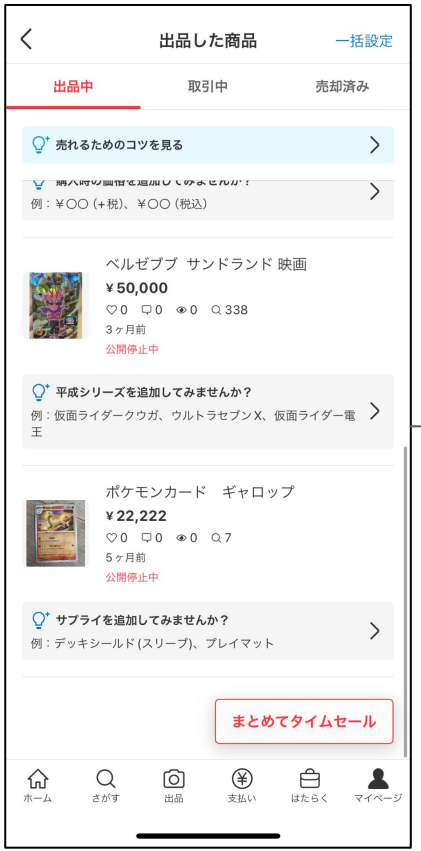
LLM Application in Action: AI Listing

Not all listings are sold-out in a timely manner for various reasons. Based on the reason, we provide different kind of **hints** to our sellers.

Utilize GPT-4 exclusively for offline extraction of listing and GPT-3.5-turbo for real-time hint.

We later tried built own LLm model. Refer to [Nejumi Leaderboard] for using best models (utilizing QLoRA to fine-tune the gemma-2b-it model):

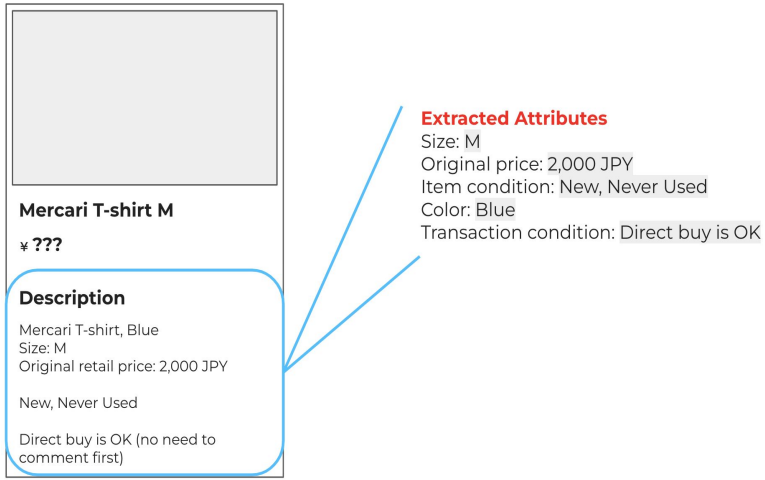
- google/gemma-2b-it · Hugging Face
- tokyotech-llm/Swallow-7b-instruct-hf · Hugging Face



Wanna add the `Series` info? e.g. デッキシールド(deck sleeve), プレイマット (playmat), ...

LLM Application in Action: AI Listing

- **Dynamic and varied attributes**: The way attributes are described can change frequently, leading to high maintenance requirements and the need for continuous model re-training.
- **New Item/category**: Large language models (LLMs) have the potential to generalize better.
- **Multi-linguality**: Most listings in Mercari are written in Japanese, however, with the huge variety of goods being exchanged, there are also listings written in other languages, such as English and Chinese



LLM Application: Similar Looks Recommendation

The Similar Looks module displays image-based recommendations similar-looking items on the item detail page

LLM Application: Similar Looks Recommendation

SigLIP is a multimodal image-text model similar to CLIP. It uses separate image and text encoders to generate representations for both modalities.

We conducted fine-tuning of the SigLIP model using approximately one million randomly sampled Mercari product listings (text-image pairs) from items listed. The input data for SigLIP consisted of **product titles (text)** and **product images (image)**, both of which were created by sellers on the Mercari platform

We use PyTorch to train and WebDataset to optimize the data loading process. Model training was conducted on a single L4 GPU via Vertex AI Custom Training. For experiment monitoring, we employed Weights & Biases (wandb).

LLM Application: Similar Looks Recommendation

SigLIP + PCA: SigLIP embeddings compressed from 768 to 128 dimensions using PCA

Offline Evaluation: user search log where user clicks are positive samples

AB Evaluation: We observed significant improvements in business KPIs (Transaction per User).

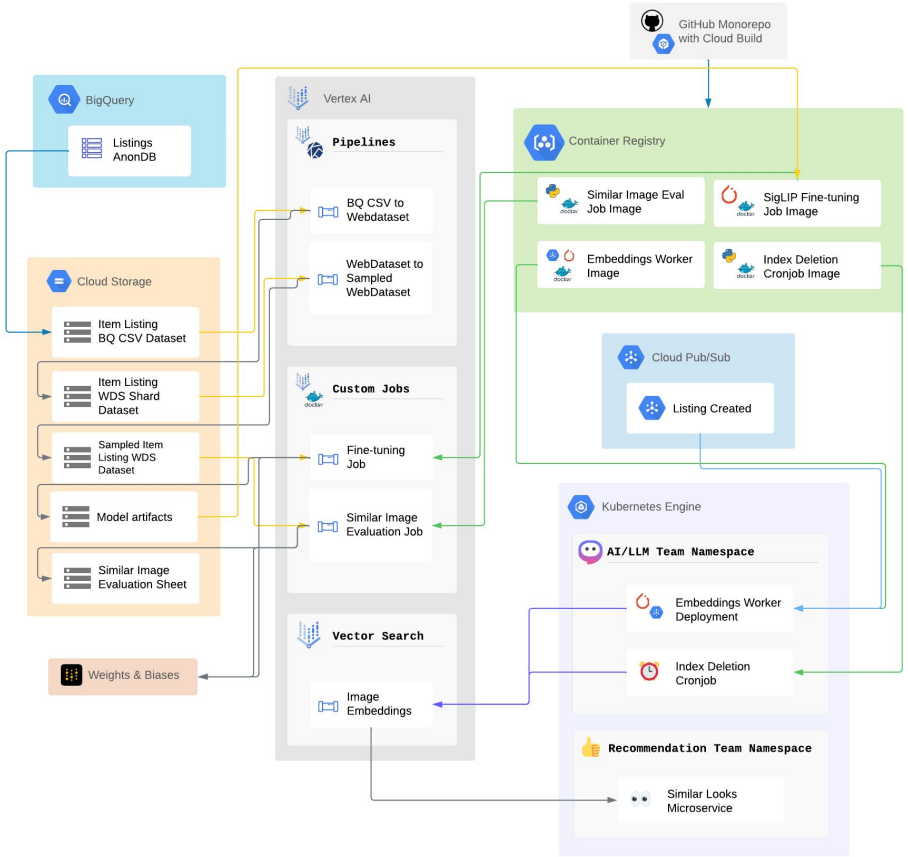
Method	nDCG@5	Precision@1	Precision@3
Random	0.525	0.256	0.501
MobileNet	0.607	0.356	0.601
SigLIP + PCA	**0.647**	**0.406**	**0.658**
SigLIP	**0.662**	**0.412**	**0.660**

mercari

LLM Application: Similar Looks Recommendation

Deployment

1. Fetches the corresponding image
2. Converts it into a fixed-length vector embedding using **SigLIP**
3. Runs PCA to reduce the dimensions for improved latency: 768 dim → 128 dim. Stores in **Vertex AI Vector Search**
4. **Inference**: The nearest-neighbor search algorithms built into Vertex AI enable fast retrieval of visually similar listings



mercari

mercari

Q&A

Kyle Liu
Head of Search&Discovery
Engineering at Mercari