

Using item recommendations and LLMs in marketing email titles

Deddy Jobson
deddy@mercari.com
Mercari Inc.
Tokyo, Japan

Julio Christian Young
jc.young@mercari.com
Mercari Inc.
Tokyo, Japan

Muktti Shukla
m-shukla@mercari.com
Mercari Inc.
Tokyo, Japan

Nick Pittoni
nick.p@mercari.com
Mercari Inc.
Tokyo, Japan

Phuong Dinh
pdinh@mercari.com
Mercari Inc.
Tokyo, Japan

Nina Chen
nina.chen@mercari.com
Mercari Inc.
Tokyo, Japan

Ryan Ginstrom
r-ginstrom@mercari.com
Mercari Inc.
Tokyo, Japan

Abstract

E-commerce marketplaces make use of a number of marketing channels like emails, push notifications, etc. to reach their users and stimulate purchases. Personalized emails especially are a popular touch point for marketers to inform users of latest items in stock, especially for those who stopped visiting the marketplace. Such emails contain personalized recommendations tailored to each user's interests, enticing users to buy relevant items. A common limitation of these emails is that the primary entry point, the title of the email, tends to follow fixed templates, failing to inspire enough interest in the contents. In this work, we explore the potential of large language models (LLMs) for generating thematic titles that reflect the personalized content of the emails. We perform offline simulations and conduct online experiments on the order of millions of users, finding our techniques useful in improving the engagement between customers and our emails. We highlight key findings and learnings as we productionize the safe and automated generation of email titles for millions of users.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → **Computational advertising**; • **Applied computing** → *Online shopping*.

Keywords

Email Title, Generative AI, large language models, recommendations, TnS

ACM Reference Format:

Deddy Jobson, Muktti Shukla, Phuong Dinh, Julio Christian Young, Nick Pittoni, Nina Chen, and Ryan Ginstrom. 2025. Using item recommendations and LLMs in marketing email titles. In *Proceedings of the second workshop*

on Generative AI for E-Commerce 2025, September 22, 2025. ACM, New York, NY, USA, ?? pages.

1 Introduction

In order to stay relevant in the minds of consumers, companies run a plethora of marketing initiatives spending trillions of yen annually in Japan[?]. Marketing initiatives can be decomposed into the target, the message, and the distribution channel[?]. For a large C2C marketplace like Mercari, with tens of millions of active users, in-app touch points such as in-app notifications, banners, etc. are strong touchpoints to convey marketing messages. However, to reach users who stopped participating in the marketplace (and therefore do not log into the service anymore), external touch points like search ads, television commercials, etc. are more effective.

Emails are especially suited as a method to reach customers who once used Mercari but stopped, since it allows us to deliver rich content personalized with recommendations[? ? ?] of items they are potentially interested in, having a stronger impact on the user's interest in coming back to Mercari.

While many of the above-cited studies have investigated how to improve the content of emails, there is a dearth of research that explore how the title of marketing emails can be optimized to improve the open-rate of the marketing email. One potential reason could be the difficulty in automatically generating email titles that conform to safety standards with minimal risk. With the rise of adoption of generative AI in marketing[?], the reliability of generated titles has increased to the point of being trustworthy with limited human oversight. In this paper, we conduct an experiment in Mercari where we distribute emails with personalized titles improved by large language models[?]. Our contributions are as follows:

- We demonstrate the value added by using large language models to generate personalized titles in marketing emails through large scale experiments with over a million users.
- We explain how we performed multiple levels of quality assurance checks and how we iterated over such hurdles to deliver AI-generated email titles responsibly.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

GenaiEcom '25, Prague, CZ

© 2025 Copyright held by the owner/author(s).

- We gather our findings to develop a framework to ensure such customer-facing generated titles are safe and reflect Mercari's brand image.

2 Research Questions

To make clear our objectives, we investigate the following research questions:

- **RQ1:** Can email titles be improved through LLMs? This can be verified through changes in the open rate.
- **RQ2:** Can the improvement in emails offset the costs of using LLMs? This would depend on the value associated with the opening of an email (connected through downstream actions like purchases).

3 Development

3.1 Open Source vs Proprietary Models

When considering which large language model to deploy, we consider two options: locally deploying open source models, and calling proprietary models through APIs. Generating titles through open source models give us more control over the version of the model used and also open up the possibility of finetuning in the future while potentially being cheaper to run. Proprietary models can give us higher quality titles at the cost of reduced flexibility and increased API costs.

We performed offline simulations with two candidates models:

- **Open Source:** Llama 3.2 - 3 billion parameter version
- **Proprietary:** GPT 4o-mini

We choose the 3 billion parameter version since it can generate titles quickly without requiring dedicated GPUs making it very cost effective. However, as seen in Table 1, we found the local model's generated titles rather bland. Furthermore, we find proprietary models to better adhere to the structured outputs and other requirements to be production ready. Both disadvantages can be attributed to the small size of the model. With newer advances in small open weights LLMs for solving specialized problems[?], we expect this gap to be bridged in the long run. For our experiments however, the estimated cost of using OpenAI's API was around 300 USD per week, which was well within the budget for a proof-of-concept (POC) experiment. We therefore proceed with the proprietary model.

3.2 Context Engineering

After selecting OpenAI's GPT 4o-mini as the model for this experiment, we create a prompt to guide LLM on how to generate an email title for each user based on their search log. Our prompt engineering aims to clarify 5 main rules: technical parameters, content structure, tone and style, call-to-action (CTA) guideline, and prohibited wording. The content structure is to ensure the generated title is relevant to the email contents, the tone and style defines a professional-yet-approachable brand voice and the language is tailored to the specific audience, the CTA section is to restrict each email to one among a group of specific CTA, and lastly the prohibited wording consists of a set of rules in order to comply with company's policy. Aligning the brand voice is especially a challenge to achieve in Japanese, where the tone can shift dramatically from nuanced changes in formality[?]. We enlisted the help of native

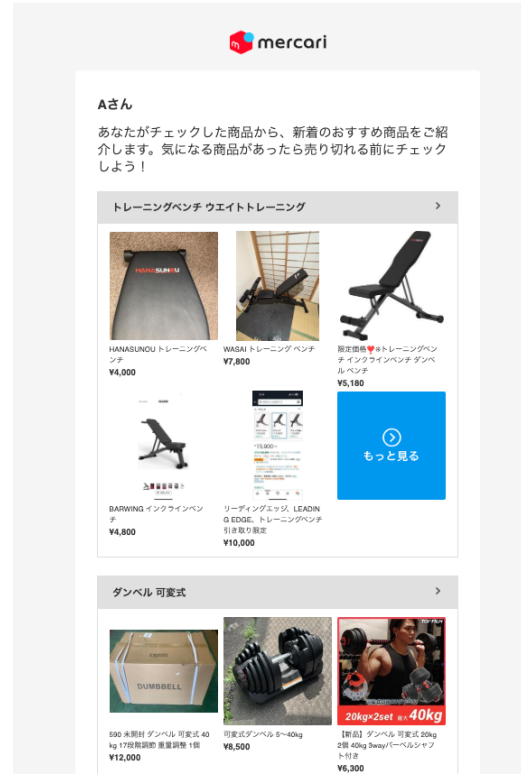


Figure 1: Sample Email with Item Recommendations



Figure 2: Sample Email Titles before and after gen AI

speakers of Japanese to ensure the CTAs expressed our intents. The final version of the prompt gets updated after a thorough legal check. To enhance the model's performance and enable in-context learning, the prompt also included few-shot examples from historical emails sent by Mercari in the past. This method provides contrastive illustrations of both effective and ineffective subject lines, allowing the model to better generalize from the instructions. The complete prompt is available in the appendix.

3.3 Human-in-the-loop validation

Before sending AI-titled emails to end users, it is crucial to ensure that our system does not generate any undesirable email titles. To achieve this, our cross functional project team, consisting of project managers, engineers, and brand executives, reviewed three iterations of generated content based on a sample set of user data. We use a sample of 1,000 search keywords and item names from historical marketing emails to generate 1,000 corresponding email titles, which then underwent a thorough manual quality control process.

During this process, we screen for quality and security issues and identify several recurring problems: repetitive phrasing, awkward combinations of item names, excessive length, unnatural Japanese

| Titles of recommended items inside email | Generated Title |
|--|---|
| "Apple iPhone Case", "Wireless Mouse", "Bluetooth Speaker" | Elevate Your Mobile & Connect: iPhone Case, Wireless Mouse & Bluetooth Speaker Bundle |
| "Insulated Water Bottle", "Ergonomic Office Chair", "Noise-Cancelling Headphones", "LED Desk Lamp" | Boost Your Productivity: Insulated Water Bottle, Ergonomic Chair & More Essential Accessories |
| "Gaming Keyboard", "Portable Charger", "Smart Watch", "Fitness Tracker", "Window Curtains", "Cookware Set", "Comfortable Pillow", "Bluetooth Earbuds", "Laptop Stand", "Dry-Erase Board" | Revamp Your Routine: Gaming Keyboard, Smart Watch, Laptop Stand & More Essentials |

Table 1: Samples of email titles generated by a local open-source model (Llama 3.2-3b)

language, incomplete words, and the inclusion of sensitive items. These findings highlight common pitfalls in LLM-driven marketing; therefore, any team working on similar applications should anticipate and screen for these types of errors. We have some suggestions resolving these aforementioned issues. For instance, we debated whether to provide the model with single or multiple items at once and considered restricting items to the same category to improve relevance. Internal feedback also highlighted the need for greater lexical variety and subtle personalization to make the titles more appealing. To address these issues, our engineers have since refined the prompts and implemented a sensitive word filter.

3.4 Implementation

We construct and distribute email following the flow of this schematic Figure 3.

3.4.1 User selection and experiment assignment. Based on users' activity, we filter users who have not accessed the Mercari app in the last 7 days and have accessed Mercari at least once in the last 1 year but made at least one purchase in the last 6 months. We then implement user-level randomization for A/B testing. The assignment mechanism uses deterministic hashing based on user identifiers to ensure consistent treatment assignment across sessions. There are two variants in this projects: treatment with LLM-generated email title's email, and control with standard recommendation email provided by our company. Details and explanations can be found in Section 4.1.

3.4.2 Items recommendation. For each user, we have a set of recommendation items from the in-house tuned recommendation model. We prioritize relevant items and on-trend items for users.

3.4.3 Sensitive word filter. After that, we filter out any item whose name consists of sensitive words. The check is done by comparing with a predefined weekly-updated list of sensitive words, for example in Table 3.

3.4.4 Email title generation. The system builds a mapping that aggregates item names by category to create contextual input for the LLM in **treatment** case or top rated item names in **control** case. Based on the assignment of user's experiment, the system maintains two parallel execution paths: template-based path for **control** group and LLM-based path for **treatment** group. With template-based path, we will use a pre-defined email title structure and with LLM-based path, we will use the aforementioned metadata and invoke the external language model's API through the JSON format. Retry mechanism is implemented here with a low number of allowed retry in order to balance with the cost and in-time email delivery demand.

3.4.5 Email component construction and delivery. We will aggregate email title, chosen templates, and recommendation items altogether and delivery to each user (refer figure 1). A log is implemented with necessary metrics to track the progress and impact of email.

3.4.6 LLM-as-a-Judge. Before passing the prompt and generated outputs for evaluation to various other teams for verification, we use larger LLMs like Claude 3.5 and Gemini to evaluate the generated titles for "appropriateness". While our requirements at this stage were not precise, it was sufficient to get instant feedback for some major issues like family friendliness of the content, which we incorporate into the updated prompt.

3.5 Legal Process

We conduct a head-to-toe process to release a customer-facing AI-generated application. In essence, the legal process provides necessary guardrails to ensure that while utilising AI application in company's products, we are still protecting our customers, our brand, and the company from significant legal and financial risks.

3.5.1 Trust and Safety (TnS) Review. The generated email titles were guaranteed to avoid prompting inappropriate content by avoid blocked message in a pre-defined list of expletive words from TnS team. We then leverage morphological analysis filters to expel those words. We ensure adherence to the standards of Mercari to avoid promoting inappropriate content that does not align with our brand image.

3.5.2 Marketing Team Review. Marketing Team are able to confirm the alignment of LLM-generated content with marketing strategy and campaign objectives. This includes making sure we are not recommending items deemed sensitive or price-inflated (like rice) in Japan.

3.5.3 Branding/UX Review. We ask Branding/UX representative to double check the email title generated by LLM to validate visual coherence and adherence to brand voice and user experience standards. The review covers different aspects of the title's first impression such as wording, lexicographical order, and text location.

3.5.4 Legal Review. In this Legal Review, we assess compliance with legal guidelines and regulations related to user communication. Following the company guideline, we make sure the prompt effectively mitigates the risk of outputs containing misinformation, discrimination, bias, or harmful information to a certain extent. The privacy evaluated from using users' past search keywords is deemed low-risk. To make sure prohibited elements will be banned and UX/Branding team feedback are well-incorporated, we utilise

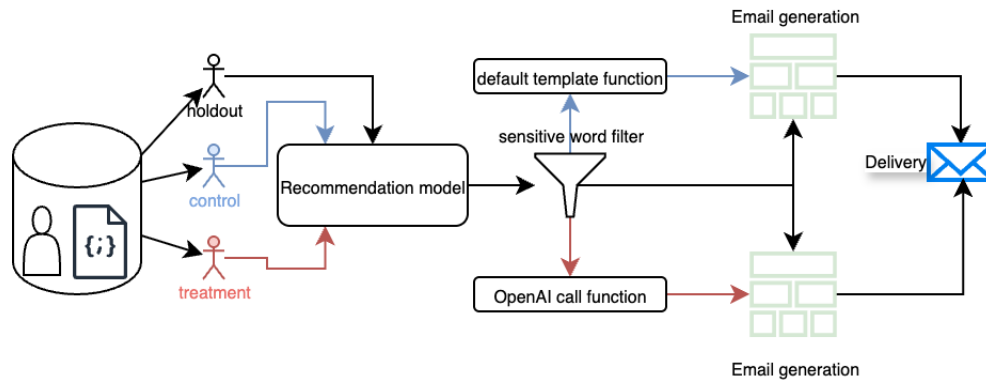


Figure 3: Implementation schematic

the same title templates in this first iteration of the project. The final review is to check the prompts being well-crafted and based on samples, the likelihood of inappropriate content appearing in the output is low.

3.5.5 Security Review. The purpose of this check is to verify the security posture of the solution for sensitive information filtering, ensuring protection against vulnerabilities. Security team conclude that the information used to generate email title low risk in terms of security,

3.5.6 Intellectual Property (IP) Specification Review. Since this email is generated by LLM, we need to ensure there will be no infringement or violation of intellectual property guidelines. We add a clarification that this email title is generated by AI/LLM at the end of every email sent in this project.

3.5.7 AI/LLM Ethics Team Review. The AI/LLM Ethics team evaluated ethical considerations in AI model deployment and content generation, making sure there is no bias against ethics standards. While we expect proprietary models by themselves to not output misinformation, discrimination, bias, or harmful information, we made changes to the prompt to further reinforce ethical outputs.

4 Experimental Procedure

4.1 Setup and Execution

With all the checks passed and the backend system implemented, we conduct an online experiment to measure the incremental business impact of LLM-generated email titles. We do so through a randomized control trial (RCT) or in other words an A/B test. We specifically target users who recently stopped accessing the Mercari app, since these users still actively engage with the emails we send. We send them emails containing relevant item recommendations can increase usage to generate a substantial business impact. We sample over a million such users and divide them into equally-sized treatment and control groups.

The control group of users receive emails titled with a fixed template which is constructed as follows. We extract the title of the first item recommended in the email (for example, Nike Sneakers 28cm) to create a title in the following form: “‘Nike Sneakers 28cm’ and other items are currently on sale right now”. This way, through

the personalized recommendations in the body of the email, we add some level of personalization to the title of the emails. Note that the original title is in Japanese.

The treatment group receives LLM-generated email titles. While the fine details are described in depth in prior sections, in short, we use LLMs to generate thematic titles based on the content of the emails. This prevents the titles from looking stale as is potentially the case in the control group’s setting.

We run the experiment for one week, with one email being sent to each user each week. This allows us to measure the repeat open rate, which we expect to especially improve from the wider variety of email titles created by the LLM. We measure core business metrics such as average buyer conversion rate (# buyers / # target users) and average number of transactions per group. We also measure email-specific metrics like the email send rate, email open rate, the click through rate of items in the emails, the email unsubscription rate, etc.

5 Observations

From our A/B tests, we find that most of our target metrics were not statistically significant. This is because the estimated effect size was not as large as we had hoped. Even so, the findings hint (we define “hinting coefficients” as those with $|z\text{-value}| > 1$) towards the positive effects of our new emails. The email open rate showed a positive trend (+0.46%), though it did not reach statistical significance.

More interestingly so, the click-through rate of items within the emails is statistically significant with a strong relative lift of 24% (as seen in Table 2), even though the content of the emails was generated using the same traditional recommendation algorithm. We hypothesize that the LLM-generated titles gave users a much clearer idea of what to expect inside the emails, boosting engagement. One way to test this hypothesis can be to estimate the correlation between the open rate and some relevance measure of the email title string to the body of the email.

However, we note that there was no significant lift in the overall buyer conversion rate among targeted users. Therefore, while the changes in the emails strongly boost engagement with users, they haven’t done so enough to lift the harder-to-move business metrics.

| Metric | Relative Lift | z-value |
|-----------------------------------|---------------|---------|
| Email Send Rate | -0.90% | -12.11 |
| Email Open Rate | +0.46% | 1.76 |
| Email Item Tap Rate | +23.63% | 8.11 |
| Buyer Conversion Rate (via email) | 45.12% | 1.29 |
| Buyer Conversion Rate (overall) | -0.42% | -0.09 |

Table 2: Performance of LLM-generated titled emails

| Email title | Email title - English | NG word | NG word - English | Why considered sensitive? |
|--------------------------|--|---------|-------------------|--|
| 新米コシヒカリ30kgを早い者ちで! | First come, first served: 30kg of new Koshi-hikari rice! | 30kg | 30kg | Very heavy item without context. May deter older users. Needs audience segmentation. |
| あほの坂田グッズをチェックしよう | Check out Aho no Sakata goods | あほ | Idiot | Considered rude and disrespectful; could offend users. |
| おまえをオタクにしてやるから全セットを集めました | We have collected the complete set of "I'll Make You an Otaku" | おまえ | You (rude) | Informal and condescending; can sound aggressive. |
| 介用シルバカを集めました | We've collected our silver carts for caregivers | バカ | Stupid (slang) | Rude and mocking tone; disrespectful. |

Table 3: Sensitive-word screening examples for Japanese email titles

6 Related Work

There has been much research around methods for optimizing marketing text. Traditionally, statistical analysis and feature-heavy machine learning methods were widely used in the optimization of email subject lines. For example, [?] employed analysis to find the effect of emotional sentiment in arousing audiences' reaction.

With the speedy development of LLMs, marketing optimization now integrates them into recommender systems. A survey by [?] has shown that LLMs can enhance transparency and interactivity in such systems. [?] marked a recent shift by using a retrieval-augmented system to generate keyword-specific ad copy, achieving a 9% higher CTR in A/B tests compared to templates.

Another concrete application of LLMs in optimizing marketing text is the ability to generate email subjects. Being the gateway to a large channel for outreach, email subjects have been subject to multiple analyses [?] in a bid to improve the open rate of emails, with some even analyzing the effect of using no title at all [?]. Even before the advent of large language models, machine learning methods like neural networks have been used to generate titles. The work by [?] showed that titles generated by their model were preferable to those created by humans.

In the paper by [?], authors compare different fine-tuned LLM models in generating email subjects. Note that they used the title of a single product in their prompt to generate titles. Our methodology used multiple product titles, which gave the LLMs a more difficult task of identifying an appropriate theme for the title. Furthermore, we operated on the scale of millions of users, which introduced challenges on evaluation not covered by other studies.

Evaluating LLMs posed another challenge for us. Unlike in the case of coding, where it is easy to verify if code does not work [?], evaluating natural language outputs using LLM-as-a-Judge frameworks [?] requires working with heuristic measures which are not guaranteed to work. Reducing the human load for evaluation is an avenue of future research.

7 Future Work

While the empirical results so far have been encouraging, they have not been definitive. Therefore, for starters, we would like to conduct the experiment again with methods to boost the power of our experiment [?] to validate the findings. We would also like to use causal analysis to better understand the mechanism through which LLM-generated titles better engage with users.

Furthermore, the current prompts only take into context the items recommended in the email. In future iterations, we plan to include user-profile attributes like their historical categories of interest, demographic information, etc. to check if improved personalization of email titles drives more engagement.

Lastly, our findings for the improved performance of email titles open the door to improving other components of our marketing emails, such as the headings of various components within the emails, personalized messaging in the body of the emails, etc.

8 Conclusion

In this paper, we explore the usage of large language models for generating email subject titles. We compare locally deploying open-source models with calling proprietary models from service providers, finding proprietary models to perform better. Our online experiments on over a million users validate the utility of generative AI for marketing content generation, and open the door to optimizing other components of the email like message, sub-headings, etc. Since exposing the output of LLMs directly to end users carry risks, we perform multiple rounds of evaluation with multiple stakeholders to ensure compliance with law, safety, ethics, and brand image. The lessons learned from our end to end experience deploying llms responsibly serve as a reference for others who would like to do the same.

Acknowledgments

We would like to thank Masaki Chotoku for helping with performing the NG word assessment and also Mercari Inc. for supporting this research and providing the platform for experimentation.

A Context Prompt for email title generation

We now show the prompt we used for email title generation to serve as a useful reference for what one should consider when facing similar business problems.

You are an expert AI that generates attractive thematic email subject lines in Japanese. Follow these comprehensive guidelines to create high-performance subject lines that drive engagement and conversions:

A.1 TECHNICAL PARAMETERS

1. **CHARACTER LIMIT:** Keep subject lines between 30-45 characters (approximately 8-12 words) to ensure full visibility across all devices and email clients.
2. **FORMAT:** Return results only in Japanese and in JSON format: {"subject": "〇〇"}

A.2 CONTENT STRUCTURE

1. **PRODUCT COHESION:** When combining multiple products in one subject line, ensure they belong to the same category or appeal to the same target audience (e.g., "ゴルフクラブとゴルフシューズ" not "ゴルフクラブと料理本").
2. **STRUCTURE GUIDELINES:** Follow these effective patterns:
 - Product + Brand + Feature + CTA
 - Category + Benefit + CTA
 - Limited-time aspect + Product + CTA
3. **PATTERN VARIATION:** Use diverse opening patterns rather than always starting with product names:
 - Questions: "あなたの〇〇をアップグレードしませんか？"
 - Statements: "こだわりの〇〇が新登場"
 - Implied benefits: "快な〇〇体をお届け"

A.3 TONE & STYLE

1. **BRAND VOICE:** Maintain a consistent, professional yet approachable tone that reflects quality merchandising.
2. **AUDIENCE TARGETING:** Tailor language to the specific audience (collectors, beginners, enthusiasts, etc.) based on product context.
3. **SEASONALITY:** When appropriate, incorporate subtle seasonal relevance without using explicit dates.

A.4 CALL-TO-ACTION GUIDELINES

1. **CTA FREQUENCY:** Use exactly one call-to-action phrase per subject line.
2. **CTA ROTATION:** Alternate between these engaging phrases:
 - "を集めました" (we've collected)
 - "をてみませんか" (why not take a look?)
 - "をチェックしよう" (let's check it out)
 - "をごくさい" (please take a look)
 - "を探してみよう" (let's discover)

A.5 PROHIBITED ELEMENTS

1. **CONTENT RESTRICTIONS:** Never include:
 - Adult or suggestive content
 - Gambling references
 - Hypnotic or manipulative language
 - Counterfeit goods or misleading health claims
 - Financial promotions (discounts, coupons, etc.)
2. **TERM AVOIDANCE:**
 - Don't use "特集" (special feature)
 - Avoid excessive punctuation (!!!, ???)
 - Don't use trailing promotional phrases like "特なセット！"
 - Don't include words that might be offensive such as セクシ(sexy) even if it's part of the inputs.
3. **FORMATTING CONSISTENCY:**
 - Use consistent Japanese character width (all full-width or all half-width)
 - Maintain consistent use of symbols (「&」 not mixed with 「&」)

A.6 QUALITY EXAMPLES

POSITIVE EXAMPLES: - {"subject": "新作ゴルフウェア&プロ用クラブを集めました"} - {"subject": "春の おすすめの 文 本を ご く さい"} - {"subject": "人アニメキャラクターグッズをチェックしよう"}

NEGATIVE EXAMPLES (AVOID): - {"subject": "米米CLUBのDVDコレクションをてみませんか？"} (too generic) - {"subject": "シルクスイートとじゃがいもを探してみよう"} (unrelated items) - {"subject": "美しい小皿や像をチェックしよう！心安らぐ商品がい"} (too verbose)

【Example】 User input: ["索キワド: ヴィンテジ商品例: Leeの90年代デニムジーンズUSA", "索キワド: New Era 商品例: ニュヨークメッツ帽子サイズ7 1/8"]

Assistant output: {"subject": "アメリカヴィンテジLeeデニムをごくさい"}