CRMAgent: A Multi-Agent LLM System for E-Commerce CRM Message Template Generation

Yinzhu Quan*
Georgia Institute of Technology
Atlanta, GA, USA
yquan9@gatech.edu

Xinrui Li
ByteDance Inc.
Seattle, WA, USA
ava.li@bytedance.com

Ying Chen
ByteDance Inc.
San Jose, CA, USA
eric.ychen@bytedance.com

Abstract

In e-commerce private-domain channels such as instant messaging and e-mail, merchants engage customers directly as part of their Customer Relationship Management (CRM) programmes to drive retention and conversion. While a few top performers excel at crafting outbound messages, most merchants struggle to write persuasive copy because they lack both expertise and scalable tools. We introduce CRMAgent, a multi-agent system built on large language models (LLMs) that generates high-quality message templates and actionable writing guidance through three complementary modes. First, group-based learning enables the agent to learn from a merchant's own top-performing messages within the same audience segment and rewrite low-performing ones. Second, retrieval-and-adaptation fetches templates that share the same audience segment and exhibit high similarity in voucher type and product category, learns their successful patterns, and adapts them to the current campaign. Third, a rule-based fallback provides a lightweight zero-shot rewrite when no suitable references are available. Extensive experiments show that CRMAgent consistently outperforms merchants' original templates, delivering significant gains in both audience-match and marketing-effectiveness metrics.

1 Introduction

Customer Relationship Management (CRM) Kumar and Reinartz [2018] messages serve as a primary lever for user retention and revenue generation on modern e-commerce platforms. Merchants routinely push promotional notifications that announce new products, highlight price drops, or deliver loyalty rewards to segmented audiences through app and messaging channels. Crafting effective templates remains a bottleneck because most merchants lack dedicated marketing expertise, work under tight time constraints, and therefore rely on generic system-generated copy. Many CRM messages consequently underperform, leaving substantial engagement and conversion potential untapped.

Recent work has shown the effectiveness of LLM agents for structured decision-making in operational domains such as logistics Ieva et al. [2025], Quan et al. [2025], supply chain management Li et al. [2023], Quan and Liu [2024b], Wang et al. [2025], and economic reasoning Quan and Liu [2024a], Guo and Yang [2024], Liu and Quan [2025a]. To address this performance gap in CRM, we present CRMAgent (Figure 1), a multi-agent system built on large language models (LLMs) that automatically improves underperforming CRM message templates in e-commerce. Using historical engagement data, CRMAgent identifies low-performing messages, retrieves high-quality examples from similar contexts, and produces improved variants tailored to each merchant and audience segment

*This work was done during a summer internship at ByteDance Inc.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

group. The workflow is divided among four specialized agents: `ContentAgent`, `RetrievalAgent`, `TemplateAgent`, and `EvaluateAgent`, each responsible for a distinct subtask.

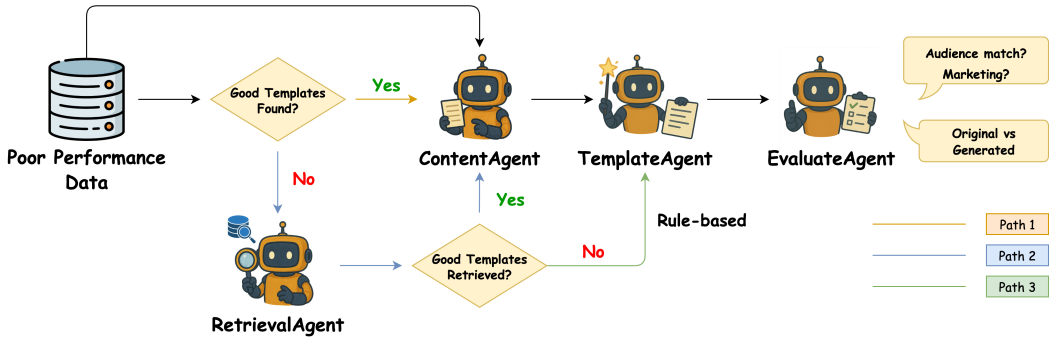

Figure 1: Framework of CRMAgent, a multi-agent system for improving underperforming CRM message templates in e-commerce.

Our contributions are threefold: (1) We present an end-to-end multi-agent LLM framework for e-commerce CRM message template generation. The system decomposes the task into four specialized agents: `ContentAgent`, `RetrievalAgent`, `TemplateAgent` and `EvaluateAgent`, which together handle data ingestion, analysis, template generation and automatic quality assessment. (2) To cover diverse data scenarios, we develop a `TemplateAgent` which can learn from a merchant's own high-performing messages within the same audience segment (group-based learning), retrieve and adapt cross-merchant exemplars that match audience segment, voucher type and product category (retrieval-and-adaptation), and fall back to a rule-based zero-shot rewrite when no suitable references exist. (3) And we implement `EvaluateAgent` which supports two evaluation modes: assigning separate scores for audience fit and marketing effectiveness to each candidate, and directly selecting the superior message.

2 Multi-Agent Collaboration for CRM Message Template Generation

This section presents the architecture of the CRMAgent system, detailing the roles and interactions of its four specialized agents.

2.1 `ContentAgent`

`ContentAgent` performs structured diagnosis of CRM message templates by comparing high- and low-performing examples within the same audience segment. It summarizes why certain messages achieved better engagement than others, focusing on structural and stylistic patterns in content design. Specifically, it analyzes differences in phrasing, organization, targeting logic, and the way incentives are presented. To ensure contextual relevance, engagement percentiles are calculated within each audience segment, defined by the plan's audience label. Figure 5 in Appendix C shows the prompt used to guide this comparative analysis.

2.2 `RetrievalAgent`

`RetrievalAgent` retrieves high-quality CRM message templates from other merchants within the same audience group, selecting those that are most similar in product category and voucher type. When no suitable high-performing templates exist within the same merchant-audience group, `RetrievalAgent` searches cross-merchant examples to support rewriting. Each candidate message is represented by a dense embedding generated using a SentenceTransformer [Reimers and Gurevych, 2019] model all-MiniLM-L6-v2 Wang et al. [2020]. Embeddings are computed based on structured metadata, including the product category hierarchy and voucher attributes. These vectors are indexed using Faiss Douze et al. [2024] for efficient similarity search. At inference time, the agent encodes the target message's metadata into a dense vector and retrieves the top-k most similar templates to serve as few-shot references for `TemplateAgent`. If both product and voucher metadata are missing, the system skips `RetrievalAgent` and falls back to rule-based rewriting.

2.3 `TemplateAgent`

`TemplateAgent` rewrites low-performing CRM message templates by combining diagnostic insights from `ContentAgent` with high-performing exemplars. These exemplars are either sourced from the same merchant and audience segment or retrieved across merchants by `RetrievalAgent`. The agent receives four types of input: the original message, a structured explanation of its weaknesses, a summary of success patterns, and one or more reference templates. To improve clarity, persuasiveness, and alignment with audience expectations while still preserving the original promotional intent such as product and voucher offers, `TemplateAgent` follows a three-stage template generation strategy: it first uses in-group exemplars if available, otherwise retrieves top-ranked examples from the same audience group with similar product and voucher characteristics, and finally generates a message using rule-based rewriting when no suitable references are found. The exemplar-guided prompt is shown in Appendix C Figure 6, and the rule-based fallback prompt is shown in Appendix C Figure 7.

2.4 `EvaluateAgent`

`EvaluateAgent` assesses the quality of CRM message template revisions through two complementary evaluation modes. In the scoring mode, it rates each message independently on two criteria: (1) audience match and (2) marketing effectiveness. Each criterion uses a three-level scale with values of 1 (poor), 3 (average), and 5 (excellent). Concise justifications are provided to support each score and enhance interpretability. In the comparison mode, the agent receives two anonymized versions labeled as A and B and selects the more persuasive and audience-appropriate candidate. The agent is not told which version is original and which is rewritten, ensuring unbiased evaluation based solely on message quality. The prompt for the scoring mode appears in Appendix C Figure 8, and the prompt for the comparison mode is shown in Appendix C Figure 9.

3 Experiments

We conduct comprehensive experiments to assess the effectiveness of our proposed multi-agent framework for CRM message template generation. This section is organized into three parts: we first present quantitative evaluation results across audience segments, analyzing audience alignment, marketing quality, and preference metrics. We then perform an ablation study to examine the impact of different LLM configurations for each agent component. Finally, we include qualitative case studies to highlight how generated messages differ from the originals in structure, tone, and targeting strategy.

3.1 Experimental Results

Table 1: Evaluation performance of original and generated CRM messages across audience segments. (Ori: Original; Gen: Generated; Pre: Preferred.)

Audience Segment	**Count**	**Audience Score**			**Market Score**			**Preference**	
		Ori	**Gen**	**Δ (%)**	**Ori**	**Gen**	**Δ (%)**	**Pre**	**Rate (%)**
Potential New Customers	701	3.32	4.71	41.87	3.14	4.78	52.23	Gen	88.16
New Buyers	681	3.63	4.38	20.66	3.32	4.57	37.65	Gen	80.62
Lapsing Buyers	515	3.84	4.12	7.29	3.41	4.38	28.45	Gen	67.18
Abandoned Cart Buyers	493	4.76	4.91	3.15	3.17	4.83	52.37	Gen	87.22
Unpaid Order Buyers	439	4.95	4.89	-1.21	3.30	4.73	43.33	Gen	83.60
Post-Purchase Group	352	4.91	4.74	-3.46	3.18	4.56	43.40	Gen	76.70
Price-Drop Group	267	4.99	4.79	-4.01	4.03	4.39	8.93	Gen	63.67
Active Old Followers	186	4.06	3.89	-4.19	3.25	4.60	41.54	Gen	70.43
Frequent Buyers	140	4.20	4.20	0.00	3.54	4.44	25.42	Gen	64.29
New Followers	125	4.70	4.46	-5.11	3.77	4.52	19.89	Gen	79.20
Repeat Buyers	58	4.17	4.00	-4.08	3.41	4.34	27.27	Gen	56.90
Overall	3957	4.18	4.56	9.09	3.33	4.61	38.44	Gen	78.44

We evaluate the quality of CRMAgent generated templates against merchant originals across 11 audience segments, including potential new customers, new buyers, lapsing buyers, abandoned cart

buyers, unpaid order buyers, post-purchase users, price-drop seekers, active old followers, frequent buyers, new followers, and repeat buyers. First, we quantify gains in audience alignment, marketing effectiveness, and overall preference using the structured scores produced by `EvaluateAgent`. Next, we compute BERTScore-F1 and chrF to verify that the rewritten copy preserves the core semantic content while allowing phrasing changes. Together, these two sets of metrics offer a comprehensive view of both persuasive impact and faithfulness to the original intent.

We use GPT-4o [OpenAI, 2024] as the underlying model for `ContentAgent`, `TemplateAgent`, and `EvaluateAgent` in this experiment. Table 1 shows that the generated templates consistently outperform the originals across most audience segments. Overall, the audience fit score rises by 9.09%, and the marketing effectiveness score jumps by 38.44%. `EvaluateAgent` chooses the rewritten version in 78.44% of blind comparisons, confirming its superiority. Qualitatively, segments that typically need a stronger push, such as potential new customers, benefit the most: the new copy adds explicit urgency and a clearer value proposition, making it more compelling to first time shoppers. In contrast, segments like frequent buyers show only modest gains because the baseline messages are already well targeted; improvements here mainly come from tighter calls to action rather than drastic content changes. A few groups, for example price-drop seekers, record a slight dip in audience fit score, likely because the rewritten text trades some price sensitivity cues for broader appeal. Even so, their marketing scores still rise, indicating that the generated messages remain more persuasive overall despite minor targeting trade-offs.

To complement the evaluation of persuasive quality, we also measure how closely the generated messages preserve the meaning and style of the originals. Table 2 reports semantic and surface-level similarity scores using BERTScore-F1 [Zhang et al., 2019] and chrF [Popović, 2015], respectively. BERTScore-F1, computed with the multilingual `xlm-roberta-large` [Conneau et al., 2019] model, captures token-level semantic alignment between the original and generated messages. Meanwhile, chrF quantifies surface-level overlap based on character and word n-gram matches. The overall BERTScore-F1 of 0.87 and chrF of 24.22 suggest that while the rewrites are not mere copies, they still maintain strong semantic consistency with the source messages—allowing for effective reformulation without drifting from the original intent.

Table 2: Similarity between original and generated CRM messages across audience segments, measured by BERTScore-F1 and chrF.

Audience Segment	**BERTScore-F1**	**chrF**
Potential New Customers	0.87	23.44
New Buyers	0.87	23.15
Lapsing Buyers	0.87	21.24
Abandoned Cart Buyers	0.87	25.26
Unpaid Order Buyers	0.87	25.30
Post-Purchase Group	0.89	30.67
Price-Drop Group	0.89	26.13
Active Old Followers	0.87	20.48
Frequent Buyers	0.86	24.49
New Followers	0.88	22.80
Repeat Buyers	0.86	22.46
Overall	0.87	24.22

4 Conclusion

In this paper, we introduce CRMAgent, a multi-agent system powered by LLMs for generating high-quality CRM message templates in the e-commerce domain. The system integrates three complementary strategies: content diagnosis, retrieval based adaptation, and rule based fallback to support merchants in crafting action oriented, persuasive copy across diverse audience segments. Extensive experiments show that CRMAgent consistently improves both audience alignment and marketing effectiveness compared to merchant written messages. In blind evaluations, the generated messages are preferred in the majority of cases, demonstrating the practical impact of our system. These findings suggest that multi-agent LLM systems can serve as scalable and effective solutions for automated CRM content creation, with potential to support merchants of varying expertise levels.

References

Yuemin Chen, Feifan Wu, Jingwei Wang, Hao Qian, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, and Meng Wang. Knowledge-augmented financial market analysis and report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1207–1217, 2024.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL `http://arxiv.org/abs/1911.02116`.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yue Guo and Yi Yang. Econnli: Evaluating large language models on economics reasoning. *arXiv preprint arXiv:2407.01212*, 2024.

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. A retrieve-and-edit framework for predicting structured outputs. *Advances in Neural Information Processing Systems*, 31, 2018.

Saverio Ieva, Ivano Bilenchi, Filippo Gramegna, Agnese Pinto, Floriano Scioscia, Michele Ruta, and Giuseppe Loseto. Enhancing last-mile logistics: Ai-driven fleet optimization, mixed reality, and large language model assistants for warehouse operations. *Sensors*, 25(9):2696, 2025.

Vineet Kumar and Werner Reinartz. *Customer relationship management*. Springer, 2018.

Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. Large language models for supply chain optimization. *arXiv preprint arXiv:2307.03875*, 2023.

Yuhang Liu, Xueyu Hu, Shengyu Zhang, Jingyuan Chen, Fan Wu, and Fei Wu. Fine-grained guidance for retrievers: Leveraging llms' feedback in retrieval-augmented generation. *arXiv preprint arXiv:2411.03957*, 2024.

Zefang Liu and Yinzhu Quan. Tpp-llm: Modeling temporal point processes by efficiently fine-tuning large language models. *arXiv preprint arXiv:2410.02062*, 2024.

Zefang Liu and Yinzhu Quan. Econwebarena: Benchmarking autonomous agents on economic tasks in realistic web environments. *arXiv preprint arXiv:2506.08136*, 2025a.

Zefang Liu and Yinzhu Quan. Retrieval of temporal event sequences from textual descriptions. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 37–49, 2025b.

Vijit Malik, Akshay Jagatap, Vinayak Puranik, and Anirban Majumder. Pearl: Preference extraction with exemplar augmentation and retrieval with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1536–1547, 2024.

OpenAI. Hello gpt-4o, May 2024. URL `https://openai.com/index/hello-gpt-4o/`.

OpenAI. Introducing gpt-4.1 in the api, April 2025a. URL `https://openai.com/index/gpt-4-1/`.

OpenAI. Introducing openai o3 and o4-mini, April 2025b. URL `https://openai.com/index/introducing-o3-and-o4-mini/`.

Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

Yinzhu Quan and Zefang Liu. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning. *arXiv preprint arXiv:2405.07938*, 2024a.

Yinzhu Quan and Zefang Liu. Invagent: A large language model based multi-agent system for inventory management in supply chains. *arXiv preprint arXiv:2407.11384*, 2024b.

Yinzhu Quan, Yujia Xu, Guanlin Chen, Frederick Benaben, and Benoit Montreuil. Leveraging large language models for risk assessment in hyperconnected logistic hub network deployment. *arXiv preprint arXiv:2503.21115*, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Harmanpreet Singh, Nikhil Verma, Yixiao Wang, Manasa Bharadwaj, Homa Fashandi, Kevin Ferreira, and Chul Lee. Personal large language model agents: A case study on tailored travel planning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 486–514, 2024.

Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. Large language model supply chain: A research agenda. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–46, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Dayu Yang, Antoine Simoulin, Xin Qian, Xiaoyi Liu, Yuwei Cao, Zhaopu Teng, and Grey Yang. Docagent: A multi-agent system for automated code documentation generation. *arXiv preprint arXiv:2504.08725*, 2025.

Mingyue Yuan, Jieshan Chen, and Aaron Quigley. Maxprototyper: A multi-agent generation system for interactive user interface prototyping. *arXiv preprint arXiv:2405.07131*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

A Related Work

Recent multi-agent LLM systems adopt modular architectures to address structured generation tasks across diverse domains. For instance, Singh et al. [2024] design role-based agents for collaborative travel planning, while DocAgent Yang et al. [2025] structures code documentation into sequential reader–writer–verifier stages. MAxPrototyper Yuan et al. [2024] coordinates theme, text, and visual elements under a central controller for UI prototyping, and InvAgent Quan and Liu [2024b] demonstrates that agent specialization benefits inventory management in logistics. Similarly, CRMAgent applies a task-specialized multi-agent framework to CRM message rewriting, with agents dedicated to content analysis, retrieval, generation, and evaluation. Complementary to modularity, retrieval-augmented generation (RAG) improves LLM output by grounding in exemplar data or structured knowledge. Retrieve-and-Edit Hashimoto et al. [2018], PEARL Malik et al. [2024], and Chen et al. [2024] leverage retrieval to enhance preference modeling and factual accuracy. Recent work also explores temporal event modeling and retrieval Liu and Quan [2024, 2025b], combining LLMs with temporal point processes to unify event semantics and timing, while Liu et al. [2024] show that fine-tuning with retrieved exemplars improves personalization. Extending these directions, CRMAgent retrieves high-performing CRM templates based on audience, product, and voucher metadata, and generates tailored messages using marketing feedback.

B Data

We begin with over 3 million successfully delivered CRM message logs from April 2025, progressively enriching each record with shop attributes, campaign metadata, product bindings, voucher information, and 7-day user engagement signals. This results in a detailed dataset indexed by plan,

user, product, voucher, and template. We then group the data by key dimensions—including merchant, template, and voucher—to compute the average engagement score and sample count for each combination. In total, we obtain 15,806 aggregated records for downstream analysis.

Figure 2 illustrates the distribution of CRM plan volume across 11 predefined audience segments, each representing a distinct stage in the customer lifecycle. For example, "Potential New Customers" and "New Buyers" account for the largest shares, suggesting merchants often focus on early-stage acquisition. In contrast, groups such as "Repeat Buyers" and "Frequent Buyers" are comparatively smaller, indicating fewer efforts are directed toward re-engagement or loyalty building. This segmentation provides essential context for interpreting downstream engagement metrics, as user behavior and responsiveness tend to vary significantly across audience types. This breakdown helps ensure that engagement patterns are interpreted within the right audience context, rather than aggregating over inherently different customer types.

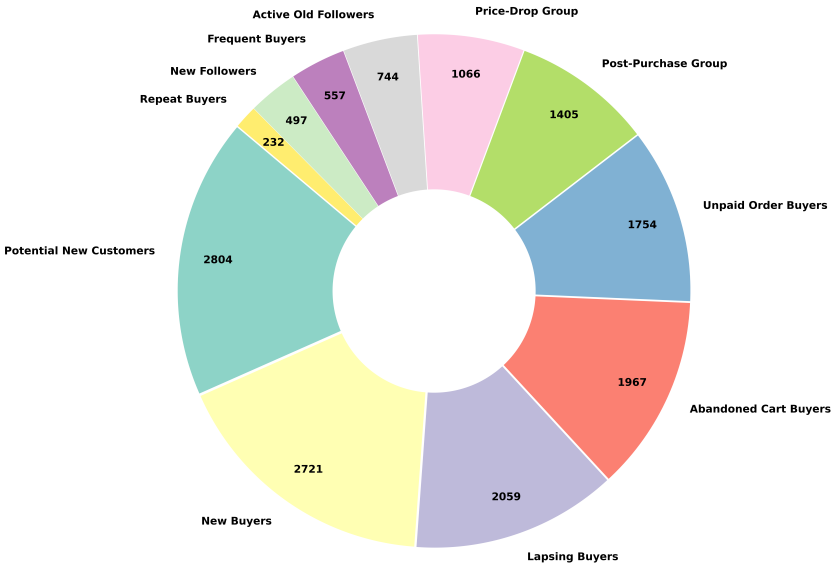

Figure 2: Distribution of CRM audience segments based on plan targeting.

The engagement score behind Figure 3 reflects weighted user behaviors within seven days of receiving a message: one point for reading, three for clicking on a store or voucher, four for clicking a product card, two for interacting with CRM buttons, and minus five for unsubscribing. We compute this score for each unique combination of merchant, template, product, audience segment, and voucher setup. The resulting distribution is highly skewed—most plans yield low engagement, while a small number achieve high scores. Since engagement levels vary significantly across audience segments (e.g., abandoned-cart users tend to be more responsive than old followers), we analyze each segment separately. Within each group, we define the top 25% of campaigns as strong performers and the bottom 25% as weak ones—these serve as contrasting examples for downstream optimization.

Figure 4 shows the distribution of CRM plans across product categories. Many plans either lack a product binding ("None") or focus on fast-moving goods like Beauty & Personal Care or Fashion Accessories. This category distribution influences not only engagement patterns but also determines how templates are generated. For instance, when no product is specified, the system falls back to rule-based generation during retrieval. In contrast, when product categories are available, we can use category-matching strategies. Thus, product type plays a central role in routing template generation.

C Agent Prompts

To enable effective collaboration among specialized agents in our system, we design dedicated prompts tailored to the role of each agent. This section presents the exact prompts used by `ContentAgent`, `TemplateAgent`, and `EvaluateAgent`, covering both scoring and comparison modes.

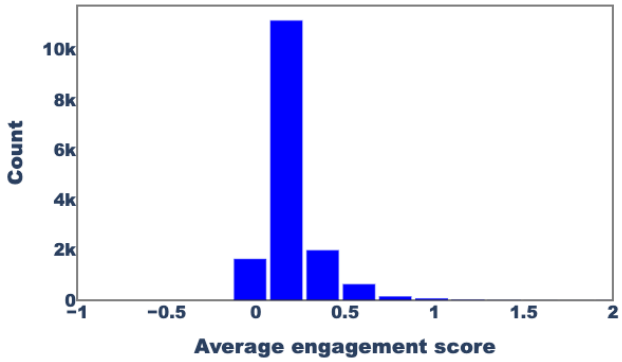

Figure 3: Distribution of average engagement scores across CRM message template plans.

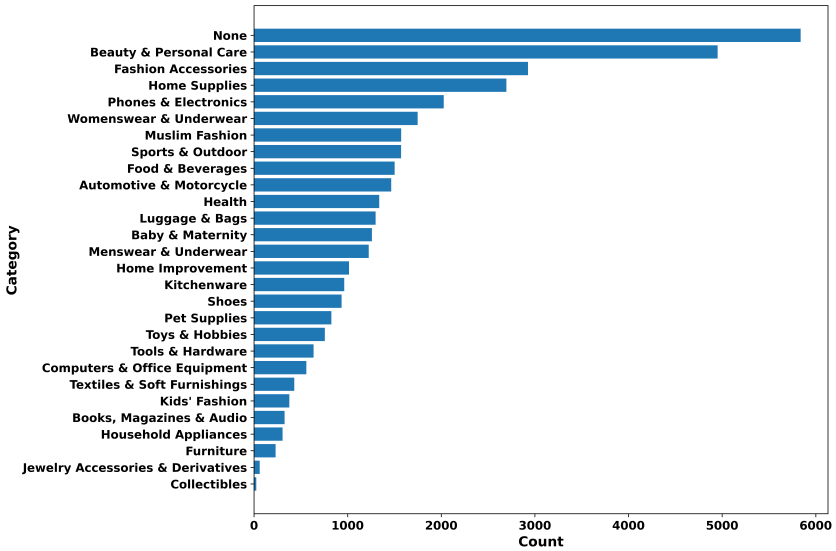

Figure 4: Distribution of CRM plans across product categories.

Prompt for `ContentAgent`

You are a CRM marketing strategist. Below are several CRM message templates sent to the audience segment `{audience}`. Each message promotes different categories and offers, with either good or poor engagement. Your task is to identify clear, actionable reasons why some templates performed better than others. Do not focus on personalization customization. Analyze and summarize the differences between high- and low-performing messages. Be concise and analytical. Focus on patterns in tone, structure, targeting, and incentive design.

=== High-Performing Templates ===
`{Template | Category | Voucher}`

=== Low-Performing Templates ===
`{Template | Category | Voucher}`

Figure 5: Prompt used by `ContentAgent` to generate comparative analyses of CRM template performance.

Prompt for `TemplateAgent`

You are a CRM marketing copywriter. A past CRM message performed poorly. Below is the Original Poor Template:
`{Poor Performance Template}`

Problem Analysis:
`{Poor Performance Reason}`

Here are some high-performing templates from the same or similar campaigns:
`{Good Performance Template}`

Summary of what makes high-performing templates successful:
`{Good Performance Reason}`

Now, please rewrite the original message into a short, action-driven, persuasive push notification. Do not focus on personalization customization. Focus on improving clarity, structure, and incentive framing — while preserving the original intent. Respond in the similar language style as the Original Poor Template.

Figure 6: Prompt used by `TemplateAgent` to rewrite underperforming CRM messages using performance diagnostics and exemplar templates.

Rule-based Prompt for `TemplateAgent`

You are a CRM copywriting assistant. Please rewrite the original message below into a short, action-driven, persuasive push notification. Do not focus on personalization customization. Focus on improving clarity, structure, and incentive framing — while preserving the original intent. Use a more engaging tone and clear incentive structure. Keep the meaning, but improve clarity and impact. Respond in the similar language style as the Original Poor Template.

=== Original Message ===
`{original}`

Figure 7: Prompt used by `TemplateAgent` for rule-based rewriting of underperforming CRM messages when no suitable exemplars are available.

Scoring Prompt for `EvaluateAgent`

You are a CRM message evaluator. You are given two marketing messages for the same audience segment, promoting the same shop, product, and voucher. The only difference is the message content. Please evaluate both Message A and Message B based on the following two criteria:
1. Audience Match — How well does the message suit the audience segment: `{audience_group}`?
2. Marketing Effectiveness — Is the message clear, persuasive, and motivating?

Use a scoring scale with three levels only: 1 = poor · 3 = average · 5 = excellent.

Audience Segment: `{audience_group}`

Message A:
`{original}`

Message B:
`{generated}`

Figure 8: Prompt used by `EvaluateAgent` to score CRM messages by audience match and marketing effectiveness.

Comparison Prompt for `EvaluateAgent`

You are a CRM message evaluator. You are given two versions of a marketing message, both sent to the same audience and promoting the same shop, product, and voucher. Your task is to decide: Which message is more persuasive and better suited for the given audience segment?

Audience Segment: `{audience_group}`

Message A:
`{original}`

Message B:
`{generated}`

Figure 9: Prompt for comparative evaluation, where `EvaluateAgent` selects the more persuasive message without knowing which version is original.

D Agent Responses

In this section, we present illustrative examples to demonstrate how `TemplateAgent` and `EvaluateAgent` operate. The `TemplateAgent` generates improved versions of underperforming CRM messages by adjusting tone, structure, and content to better align with audience expectations. The `EvaluateAgent` then compares the original and rewritten messages, scoring each in terms of audience match and marketing effectiveness, and selecting the preferred one.

D.1 `TemplateAgent` Responses

To better understand how the `TemplateAgent` improves CRM messaging, we rewrite all message templates associated with poor-performing campaigns, as identified within each audience segment. Figures 10 and 11 illustrate representative examples of these rewrites for two specific audience types.

Figure 10 targets potential new customers who have not yet established a strong relationship with the brand and often require clear incentives to engage. The original message uses a soft, emotional tone ("we really missed you") that, while friendly, lacks specificity and urgency. Such messaging may fail to capture the attention of users who are still undecided or unaware of the brand's value proposition. In contrast, the rewritten version is more strategic: it introduces urgency with phrases like "limited-time steals" and emphasizes concrete benefits such as "guaranteed savings". This not only strengthens the call to action but also aligns with the psychological triggers that are more effective for acquisition-focused messaging. The revised copy is concise, benefit-driven, and better tailored to convert first-time buyers by highlighting what they stand to gain immediately.

GPT-4o Response by `TemplateAgent` for Potential New Customers

Original Title: Check out our bestsellers!
Original Body: It's been a while since we last saw you, we really missed you and we want you to know about our ongoing special deals!

Generated Title: Limited-Time Steals on Top Picks!
Generated Body: Discover our bestsellers back in stock! Hurry and grab these limited-time deals before they're gone. Don't miss out on guaranteed savings!

Figure 10: Illustrative example of `TemplateAgent` response for the potential new customer audience segment.

Figure 11 showcases a message tailored to active old followers, a segment that already has a relationship with the brand and tends to be more responsive to loyalty-oriented communication. The original message offers a vague incentive by mentioning "a special offer for your first order" and closes with a generic "shop now" prompt, which may feel impersonal or insufficient for users

who have previously engaged. In contrast, the rewritten message opens with an explicit expression of appreciation and provides a clear benefit—twenty percent off the next order. This wording reinforces the value of continued engagement while making the reward tangible. The closing line, "before this offer disappears," introduces urgency in a natural way, encouraging immediate action without sounding overly pushy. Overall, the revised message better aligns with the expectations and motivations of returning followers by combining gratitude, specificity, and timing.

GPT-4o Response by `TemplateAgent` for Active Old Followers

Original Title: You followed – now treat yourself!
Original Body: As a thank you for following, here's a special offer for your first order. Don't miss out—shop now!

Generated Title: Exclusive Deal Just for You—Limited Time!
Generated Body: We appreciate your support! Enjoy 20% off your next order—shop now before this offer disappears!

Figure 11: Illustrative example of `TemplateAgent` response for the active old followers audience segment.

D.2 `EvaluateAgent` Responses

To prevent agents from being biased by labels such as "original" or "generated", each pair of CRM messages is anonymized as Message A and Message B during evaluation. While A and B correspond to the original and generated versions respectively, the evaluators are unaware of which is which—ensuring judgments focus solely on content quality.

For the potential new customer segment, Figure 12 illustrates a clear advantage for Message B in both audience relevance and marketing effectiveness. Message A assumes a prior relationship ("we really missed you"), which is inappropriate for new users. In contrast, Message B makes no such assumption and instead uses urgency ("limited availability") and value ("guaranteed savings") to attract first-time buyers. As a result, Message B receives the highest possible scores for both alignment and persuasion.

GPT-4o Response by `EvaluateAgent` for Potential New Customers

Audience Match Score A: 1
Audience Match Reason A: Message A refers to seeing the audience again and missing them, which implies a prior relationship. This is not suitable for 'potential new customers' who have no prior engagement history.

Marketing Score A: 3
Marketing Reason A: The message highlights bestsellers and special deals, which can be persuasive. However, the repetition of phrases and lack of urgency may reduce its overall effectiveness.

Audience Match Score B: 5
Audience Match Reason B: Message B is well-suited for 'potential new customers' as it does not imply any previous interaction and focuses on attraction through new opportunities.

Marketing Score B: 5
Marketing Reason B: Message B efficiently communicates urgency and limited availability, both of which are clear and persuasive marketing strategies that motivate immediate action.

Figure 12: Illustrative example of `EvaluateAgent` response (audience match and marketing scores) for the potential new customer audience segment.

Similarly, in the active old followers segment shown in Figure 13, Message A misfires by offering a "first order" discount to users who are likely repeat customers. This mismatch undermines both

credibility and clarity. Message B, on the other hand, acknowledges prior engagement and offers a concrete benefit ("20% off your next order"), paired with strong calls to action. It better suits the audience profile and earns top scores for both audience fit and marketing strength.

GPT-4o Response by `EvaluateAgent` for Active Old Followers

Audience Match Score A: 1
Audience Match Reason A: The message references a 'first order,' which does not align with the 'crm_Active_Old_Followers' segment, as these are existing followers likely already familiar or customers, making the offer irrelevant or confusing.

Marketing Score A: 3
Marketing Reason A: The call to action is unclear and potentially misleading since it references a 'first order' for a segment that likely already made orders or interacted previously. Lack of a specific discount or clear benefit reduces persuasiveness and motivation.

Audience Match Score B: 5
Audience Match Reason B: The message suits active old followers by acknowledging their ongoing support and offering a meaningful, clearly stated discount for their next order, relevant to established customers.

Marketing Score B: 5
Marketing Reason B: The message is clear, concise, and motivating with a strong call to action and a specific 20% discount, creating urgency with 'limited time,' making it persuasive and effective.

Figure 13: Illustrative example of `EvaluateAgent` response (audience match and marketing scores) for the active old customer audience segment.

For the potential new customer segment, Figure 14 shows that `EvaluateAgent` strongly prefers Message B. This message avoids assuming any prior customer relationship and instead focuses on driving action through urgency ("limited-time steals") and clarity ("guaranteed savings"). These are precisely the types of persuasive techniques that appeal to new customers who may be unfamiliar with the brand and require strong incentives to convert. In contrast, Message A's tone—"we really missed you"—is clearly targeted at returning users and thus mismatched for this audience. The agent correctly identifies this discrepancy and selects the more appropriate and effective version.

GPT-4o Response by `EvaluateAgent` for Potential New Customers

Preferred Message: B
Reason: Message B is more persuasive and better suited for the 'crm_potential_new_customers' segment because it creates a sense of urgency with phrases like 'Limited-Time Steals' and 'Hurry and grab these limited-time deals before they're gone.' It emphasizes scarcity and immediate action, which are effective in motivating new customers to make a purchase. Additionally, the message is concise, clear, and highlights guaranteed savings, which appeals strongly to potential new customers looking for value. In contrast, Message A feels more like a re-engagement message for existing customers ('It's been a while since we last saw you'), which is less relevant for potentials who may not have interacted with the shop before.

Figure 14: Illustrative example of `EvaluateAgent` preference decision for the potential new customer audience segment.

In the case of active old followers (Figure 15), `EvaluateAgent` again prefers Message B for its audience fit and marketing clarity. Message B directly acknowledges the customer's continued support and provides a well-specified incentive—"20% off your next order"—that feels both personal and actionable. Meanwhile, Message A refers vaguely to a "first order", which can confuse or alienate users who are already loyal. The agent's decision highlights the importance of aligning promotional language with customer lifecycle stage—what works for new users may backfire with returning ones.

GPT-4o Response by `EvaluateAgent` for Active Old Followers

Preferred Message: B
Reason: Message B is more persuasive and better suited for the 'crm_Active_Old_Followers' segment because it emphasizes exclusivity and urgency with phrases like 'Exclusive Deal Just for You' and 'Limited Time,' which can motivate immediate action. It also offers a clear, concrete discount (20% off your next order), which is more compelling than the vague 'special offer for your first order' in Message A. Since the audience consists of active old followers, highlighting a discount on the next order (rather than a first order) acknowledges their prior engagement, making the message feel more personalized and relevant.

Figure 15: Illustrative example of `EvaluateAgent` preference decision for the active old customer audience segment.

E Template Quality Analyses

We conduct a systematic analysis of CRM message template quality using our custom evaluation sheet (Appendix F). Each template is assessed by o3 [OpenAI, 2025b] and assigned exactly one error type according to predefined criteria. Figure 16 presents the distribution of identified errors.

We find that the most common issue is Weak Call-to-Action (2,057 cases), where messages fail to provide a clear or urgent prompt for user action. Vague Incentive appears frequently as well (1,290 cases), often reflecting a lack of concrete or detailed value propositions. Misaligned Tone accounts for 570 cases, indicating mismatches in emotional or stylistic alignment between message and audience. Audience Assumption Errors and Irrelevant Offers occur less frequently but still signal problematic assumptions about customer history or behavior. This motivates the development of `TemplateAgent`, which is guided by these insights to generate templates that better match audience needs and communication goals.

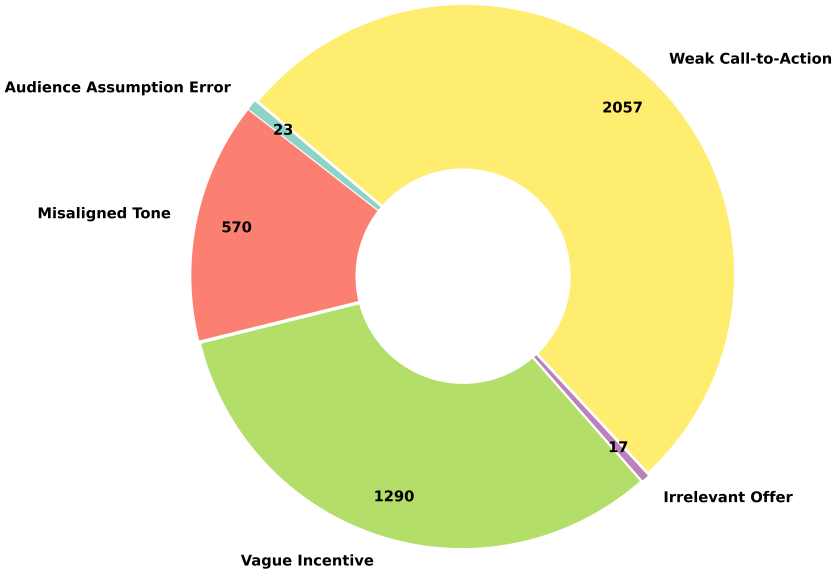

Figure 16: Distribution of quality issues in original CRM templates based on automated evaluation using o3.

F Instruction Sheet for Template Review

Provided Materials

- Original Template: Each example includes only the original poor performance CRM message template.
- Audience Segment: The target customer segment for the message is specified (e.g., potential new customers, active old followers).
- Voucher: If applicable, the corresponding voucher information is provided (e.g., voucher title, discount type).
- Product: If applicable, the corresponding promoted product information is included (e.g., product name or category).

Steps to Follow

- Read the message carefully.
- Categorize the problem using the error types below.
- Provide short feedback explaining your choice.

Error Types and Criteria

- Audience Assumption Error
  - Description: Assumes incorrect prior engagement with the user.
  - Indicators: References to past interaction for new customers; offers for first-time purchase to repeat customers.
  - Example: "Thanks for following!" sent to non-followers.
- Weak Call-to-Action
  - Description: Lacks clarity or urgency in motivating the user to act.
  - Indicators: Generic phrases; no time-sensitive or action-oriented language.
  - Example: "Shop now" with no clear reason or benefit.
- Vague Incentive
  - Description: Mentions a promotion without specifying what it is.
  - Indicators: Use of terms like "special offer" with no details on discount or benefit.
  - Example: "Here's a deal for you!" without content.
- Misaligned Tone
  - Description: The tone does not match the audience's familiarity level.
  - Indicators: Overly casual or formal; assumes emotional connection without basis.
  - Example: "We miss you!" to someone who never interacted.
- Irrelevant Offer
  - Description: The offer does not fit the customer's likely status or behavior.
  - Indicators: Offering first-order discounts to frequent buyers.
  - Example: "Your first order!" for loyal customers.

Example Application

- Audience Segment: Potential New Customer
  - Message: "It's been a while—check out our bestsellers!"
  - Identified Error(s): Audience Assumption Error
  - Feedback: This implies a prior relationship which does not exist. Rewrite to welcome the new visitor and highlight value.
- Audience Segment: Active Old Follower
  - Message: "Thanks for following—enjoy your first-order discount!"
  - Identified Error(s): Irrelevant Offer, Vague Incentive
  - Feedback: This audience is not new; avoid 'first order' wording. Clearly state what the benefit is.

G Ablation Studies

To assess how different LLM configurations affect system performance, we conduct an ablation study by varying the backbone models used in `TemplateAgent` and `EvaluateAgent`. Specifically, we explore combinations involving GPT-4o [OpenAI, 2024], GPT-4.1 mini [OpenAI, 2025a], and DeepSeek-R1 [Guo et al., 2025], as well as the use of chain-of-thought (CoT) [Wei et al., 2022] prompting. `ContentAgent` is kept fixed as GPT-4o to isolate the effects of template generation and evaluation components.

Table 3 shows that while the fully GPT-4o configuration delivers strong overall results, it is not uniformly superior. For example, applying chain-of-thought prompting to `EvaluateAgent` slightly boosts preference selection compared to the baseline, suggesting benefits in evaluation reasoning. In contrast, using CoT in `TemplateAgent` leads to degraded outcomes, likely due to verbose or unfocused generation. Models like GPT-4.1 mini offer performance comparable to GPT-4o, achieving similarly strong improvements in both audience and market scores, as well as high preference rates. This suggests that lighter-weight models can serve as viable drop-in replacements when resource constraints are a concern. In contrast, DeepSeek-R1 as `EvaluateAgent` yields the highest score improvements, but its relatively low preference rate indicates a potential mismatch between its scoring criteria and human-aligned quality judgments.

Table 3: Ablation study comparing different LLM configurations for each agent module. (Ori.: Original; Gen.: Generated; Pre.: Preferred.)

Content	**Template**	**Evaluate**	**Audience Score**			**Market Score**			**Preference**	
			Ori.	**Gen.**	**Δ (%)**	**Ori.**	**Gen.**	**Δ (%)**	**Pre.**	**Rate (%)**
GPT-4o	GPT-4o	GPT-4o	4.18	4.56	9.09	3.33	4.61	38.44	Gen	78.44
GPT-4o	GPT-4o + CoT	GPT-4o	4.27	4.26	-0.23	3.45	4.44	28.70	Gen	61.51
GPT-4o	GPT-4o	GPT-4o + CoT	3.80	4.30	13.16	3.25	4.49	38.15	Gen	81.17
GPT-4o	GPT-4.1 mini	GPT-4o	4.19	4.54	8.35	3.32	4.60	38.55	Gen	78.47
GPT-4o	GPT-4o	GPT-4.1 mini	4.20	4.43	5.48	3.30	4.61	39.70	Gen	81.05
GPT-4o	GPT-4o	DeepSeek-R1	3.39	4.51	33.04	3.08	4.84	57.14	Gen	55.40

H Case Studies

To illustrate how the CRMAgent adapts its messaging strategy across audience types, Table 4 presents two representative examples from potential new customers and abandoned cart buyers segments. In the first case, the original message is generic and celebratory, while the generated version introduces urgency ("Limited Time Offer") and a clearer call to action, better aligning with the motivational needs of first-time shoppers. In the second example, the original offers a mild reminder, but the revised message adds a direct discount ("RM3 off") and an upsell condition, using persuasive language to prompt checkout completion. These cases highlight how the agent enhances promotional effectiveness through more targeted and action-oriented rewrites.

Table 4: Illustrative case study comparing original and generated CRM message templates for the potential new customers and abandoned cart buyers segments.

Case Study 1: Potential New Customers
Original: Everyone's loving these! Shoppers love these best-sellers—find your next favorite in our latest collection. **Generated:** Limited Time Offer! Don't miss out on our hottest best-sellers! Explore our latest collection now and be the first to secure your favorites before they're gone. Shop today!
Case Study 2: Abandoned Cart Buyers
Original: Did you forget something? Looks like you left something in your cart. Ready to pick up where you left off? **Generated:** Your cart is waiting—grab RM3 off now! You left great picks in your cart! Complete your purchase with RM3 off when you spend RM15 or more. Don't miss out—shop now and enjoy the benefits!