LLM-IR: Leveraging Large Language Models for Intent Recognition in Multimodal Dialogue Systems

Junyi Wang
School of Economics and Management
Tsinghua University
Beijing, China 100084
junyi-wa24@mails.tsinghua.edu.cn

Yuanpei Sui
School of Economics and Management
Tsinghua University
Beijing, China 100084
suiyp24@mails.tsinghua.edu.cn

Tao Liu
Department of Computer Science and Technology
Tsinghua University
Beijing, China 100084
sxtegg2007@126.com

Abstract

This research tackles the complex challenge of intent recognition in multimodal dialogue systems by introducing a novel approach that leverages large language models (LLMs). By fine-tuning a state-of-the-art model using Low-Rank Adaptation (LoRA), we achieve significant performance improvements. To address the limitations of traditional methods, we employ a suite of advanced augmentation techniques, including Optical Character Recognition (OCR) for text extraction, along with image cropping, rotation, color adjustments, and text transformations such as synonym replacement and syntactic reordering. Additionally, we integrate knowledge distillation and Retrieval-Augmented Generation (RAG) techniques to incorporate external knowledge, further boosting the model's performance. Through comprehensive ablation studies and meticulous parameter tuning, our model surpasses the baseline performance by 5.35%, demonstrating the substantial benefits of utilizing LLMs in multimodal intent recognition.

1 Introduction

In the current e-commerce landscape, user intent recognition has become increasingly critical. The core competitiveness of e-commerce platforms relies not only on the variety and pricing of products but also on the ability to accurately understand user needs and respond promptly. Multimodal dialogue systems can comprehensively capture user intent by integrating multiple input modalities such as text, voice, and images, thereby enhancing user experience and enabling applications like precise recommendations and personalized marketing[1]. For example, when browsing products, users may inquire via voice, send images, or use specific expressions to convey their needs. By effectively fusing and analyzing information from these diverse modalities, the system's response speed and accuracy can be significantly improved, providing more relevant recommendations and assistance to users. This capability is especially important in e-commerce, where understanding potential purchase intent can increase conversion rates and improve user retention.

Existing research has made significant strides in this area. Multimodal fusion techniques help reduce the risk of information loss or misinterpretation by integrating data from different modalities. Deep learning-based multimodal network architectures, such as Transformers and BERT, have been widely applied in multimodal dialogue systems, enabling the capture of richer user intents at the semantic

level. Recent advancements in deep multimodal learning have been extensively used in information fusion, semantic understanding, and other areas. For example, Chen et al. reviews the current research status and challenges of multimodal dialogue systems, focusing on how to improve intent recognition accuracy through fusion of information from different modalities[1]. Ni et al. discusses deep learning-based multimodal dialogue system architectures, highlighting methods and challenges in multimodal information fusion and comparing various technical approaches[6]. Moreover, Ramachandram and Taylor examines how effective fusion of text, voice, and images in e-commerce scenarios can enhance user intent recognition accuracy and system response speed[7].

However, effectively integrating information from heterogeneous modalities (such as text, images, and voice) remains a significant challenge in multimodal dialogue systems. Many large multimodal models suffer from intent misclassification and omission of crucial information, especially when dealing with complex scenarios, leading to reduced user satisfaction and impacting sales performance on e-commerce platforms. Thus, improving the accuracy and efficiency of multimodal intent recognition has become a crucial area of technological innovation in the e-commerce sector.

2 Motivation

In the field of intelligent interaction, intent recognition has become one of the core technologies in multimodal dialogue systems, especially for e-commerce customer service and smart assistant applications. Accurately understanding user intent is vital for enhancing user experience and optimizing system responses. Despite the excellent performance of large general-purpose models in multitask learning and their support of extensive knowledge bases, their effectiveness in handling domain-specific tasks—especially in complex e-commerce dialogue scenarios—remains limited. While these models can provide broad services in open domains, they often struggle with inaccurate intent recognition, information omission, and poor context understanding when faced with specialized, scenario-specific tasks. These shortcomings negatively affect system response efficiency and user satisfaction, which are critical in the e-commerce context. Therefore, fine-tuning general models to adapt to the specific needs of the e-commerce domain has become a central motivation for the current technological breakthroughs.

In e-commerce customer service dialogue scenarios, the primary challenge in user intent recognition lies in the diversity of user queries and needs. Although users often inquire about product information, their modes of expression—such as text, voice, or images—can vary significantly. These variations create substantial differences in how user needs are expressed. Traditional large general-purpose models often struggle to efficiently parse these complex, domain-specific user intents, resulting in decreased accuracy in intent recognition. Therefore, fine-tuning large models to address domain-specific needs not only enhances their performance in e-commerce environments but also optimizes recommendation algorithms, improves customer satisfaction, and increases conversion rates. By using e-commerce-specific corpora and multimodal data for targeted training, models can better capture the unique language patterns and user behaviors inherent in the e-commerce domain, thereby significantly enhancing platform intelligence and fostering business model innovations.

From a technological perspective, overcoming the bottleneck of intent recognition not only advances the practical application of dialogue systems but also promotes interdisciplinary innovation across fields such as natural language processing, computer vision, and others. Recent research has shown that domain-adaptive fine-tuning of large models can substantially improve intent recognition accuracy, particularly in the joint analysis of multimodal data. Thus, building precise and efficient intent recognition systems by effectively integrating multimodal information, such as text, voice, and images, in e-commerce scenarios has become a key research direction in the field of intelligent interaction.

3 Data Description

This study utilizes a dataset designed for multimodal classification tasks, which consists of both textual and image data. The dataset is divided into two primary categories: Image Scene Classification and Dialogue Intent Classification.

3.1 Image Scene Classification

The image scene classification task involves classifying images sent by users to customer service into various predefined e-commerce scenarios. The following labels are used for classification:

- **Product Category Options**: Images showing product color, size, or specification options.
- **Product Main Image**: The main product image displayed on an e-commerce platform.
- **Product Detail Page Screenshot**: Screenshots from various parts of a product detail page.
- **Order Error Page**: Images showing an error, such as a "purchase failure" window during checkout.
- **Order Details Page**: A screenshot of the order details page showing complete order information.
- **Payment Page**: Screenshots showing payment options and payment success.
- **Review Page Screenshot**: Screenshots from the review section of an e-commerce platform.
- **Logistics Page - List View**: A screenshot showing a list of logistics information.
- **Logistics Page - Tracking View**: Screenshots showing the logistics path and tracking information.
- **Logistics Page - Error View**: A screenshot showing logistic error messages.
- **Refund Page**: Screenshots showing refund information.
- **Return Page**: Screenshots related to product returns.
- **Exchange Page**: Screenshots showing the exchange process.
- **Shopping Cart Page**: A screenshot of the shopping cart, including the item list and total amount.
- **Storefront Page**: A screenshot of the e-commerce store's homepage.
- **Promotion Page**: A screenshot showing special offers or discounts.
- **Coupon Page**: Screenshots showing how to claim coupons.
- **Account Page**: Screenshots showing transaction details, asset lists, or coupon information.
- **Complaint/Report Page**: Screenshots showing the complaint or report page.
- **Physical Photos (including After-Sales)**: User-uploaded photos of physical items, including damage or missing items for after-sales requests.
- **External App Screenshots**: Screenshots from third-party apps (e.g., JD.com, Pinduoduo, etc.).
- **Platform Intervention Page**: Screenshots showing customer service intervention by the platform.
- **Other Categories**: Other images that don't fit into the above categories.

3.2 Dialogue Intent Classification

The dialogue intent classification task involves determining the user's intent based on the history of the conversation and the current query. The dialogue history includes at least one image sent by the user that can help determine the intent. The intent labels are as follows:

- **Feedback on Poor Sealing**: Users report issues with the product's sealing.
- **Is it Easy to Use?**: Users ask about the usability of the product.
- **Will it Rust?**: Users inquire whether the product will rust.
- **Drainage Method**: Users ask about the drainage method for certain appliances like washing machines or water heaters.
- **Packaging Difference**: Users ask about the differences between product packaging.
- **Shipping Quantity**: Users ask about the number of items being shipped.

- **Post-Use Symptoms**: Users report symptoms after using the product.
- **Material of the Product**: Users inquire about the material used in the product.
- **Effectiveness/Function**: Users ask about the function or effectiveness of the product.
- **Fading Resistance**: Users inquire if the product is prone to fading.
- **Applicable Season**: Users ask about which season the product is suitable for.
- **Adjustable Brightness**: Users ask if the product allows adjustment of brightness, light source, or color temperature.
- **Model/Version Difference**: Users inquire about differences between two versions or models of a product.
- **Single Item Recommendation**: Users ask for recommendations on a particular product.
- **Usage Instructions**: Users ask for the method and dosage for using the product.
- **Control Methods**: Users inquire about how to control the product, including remote control options.
- **Release Date**: Users ask about when the product was released.
- **Product Specifications**: Users inquire about the product's specifications such as quantity, weight, capacity, etc.
- **Signal Strength**: Users inquire about the signal strength of mobile devices.
- **Maintenance Methods**: Users ask how to maintain the product.
- **Set Recommendations**: Users ask for recommendations for a product set.
- **Restocking Time**: Users inquire about when a product will be restocked.
- **Bubble Issues**: Users ask how to avoid or remove bubbles when applying a film to a product.

3.3 Dataset Access

The dataset includes:

- 1,000 labeled dialogue samples for training.
- 10,000 unlabeled samples for preliminary round testing.
- An additional 10,000 new test samples for the final round.

3.4 Evaluation Metrics

The performance of the model will be evaluated using the weighted F1 score across all classification labels in the test set. The weighted F1 score is calculated as:

$$WeightedAverageF1 = \frac{\sum_{i=1}^{N} w_i \times F1_i}{\sum_{i=1}^{N} w_i}$$

Where $F1_i$ represents the F1 score for class i, and w_i is the weight for each class, reflecting the class distribution.

3.5 Data Example

The following examples illustrate the task of multimodal classification based on user and customer service dialogues, including both text and image classification. These examples show how user queries and images are classified into predefined categories.

Example 1 User and Customer Service Dialogue:

> User: `<image>`
> Customer Service: Sorry, currently we can't recognize images. Could you please describe the problem briefly?
> User: `<http>`

Customer Service: If you have further questions, reply “” for additional assistance.
User: Is the inner tank made of 316 stainless steel?
Customer Service: The upgraded antibacterial stainless steel is improved based on 304 stainless steel and has antibacterial properties. However, its acid and corrosion resistance is not as good as 316 stainless steel. 316 stainless steel performs better in terms of corrosion and acid resistance but lacks the antibacterial function.

Classification Label: *Product Material*

Image:

Figure 1: Products Display Page.

Example 2 User and Customer Service Dialogue:

User: <http>
Customer Service: Isolation series — Capacity: 35g. Suitable for different skin types.

User: <image>
Customer Service: Dear user, we have reviewed your message. If you have any further questions, feel free to let us know.
User: Is it for one or two units?

Classification Label: *Shipping Quantity*

Image:

Figure 2: Product Selection Details Page.

Example 3 User and Customer Service Dialogue:

Picture 1: <image>
You are an expert in identifying images for e-commerce. Please classify the uploaded image. You only need to provide the classification result, without additional commentary. The available classification labels are: ["Physical Photos (including After-Sales)", "Product Category Options", "Product Main Image", "Product Detail Page Screenshot", "Order

```
Error Page", "Order Details Page", "Payment Page", "Customer
Service Chat Page", "Review Page Screenshot", "Logistics Page
- List View", "Logistics Page - Tracking View", "Logistics
Page - Error View", "Refund Page", "Return Page", "Exchange
Page", "Shopping Cart Page", "Storefront Page", "Promotion
Page", "Coupon Page", "Account Page", "Personal Information
Page", "Complaint/Report Page", "Platform Intervention Page",
"External App Screenshots", "Other Category Images"].
```

Classification Label: *Logistics Page - Tracking View*

Image:

Figure 3: Logistics Details Page.

4 Methodology

This study addresses the challenge of intent recognition in multimodal dialogue systems by proposing a novel approach that leverages large language models (LLMs) to enhance recognition performance. By fine-tuning an advanced framework using Low-Rank Adaptation (LoRA), we significantly improve model performance[10]. To overcome the limitations of traditional methods, we employ a variety of data augmentation techniques, including OCR extraction, image cropping, rotation, color adjustments, and text-based methods such as synonym replacement and syntactic reordering. Additionally, we integrate cutting-edge techniques like knowledge distillation and Retrieval-Augmented Generation (RAG) with large language models, incorporating external knowledge bases for further performance enhancement. Through systematic ablation experiments and careful parameter tuning, our model outperforms baseline models by 5.35%, demonstrating that leveraging large language models can achieve significant advances in multimodal intent recognition.

4.1 Large Language Models (LLMs) and LoRA Fine-Tuning

The core method of this study is the use of large language models (LLMs) for intent recognition. These LLMs (such as Qwen2-VL) perform excellently on natural language tasks and can effectively perform multitask learning. However, although these general models excel across many tasks, their performance in specific domains (like e-commerce) is often limited. To enhance model performance on domain-specific tasks, we apply LoRA (Low-Rank Adaptation) technology, which fine-tunes model weights to adapt the model to the specific needs of the e-commerce domain without requiring a full retraining of the model [10].

4.2 Data Augmentation Techniques

To improve the model's adaptability to diverse inputs, we utilize various data augmentation techniques to simulate different input variations. This approach improves the model's robustness, particularly in handling multimodal inputs [2]. Specifically, we employ the following methods:

1. **OCR Extraction**: In e-commerce dialogues, users may upload images containing product information or query content. Using Optical Character Recognition (OCR), we extract text from these images and input it into the model. This ensures that key information in images is fully utilized to enhance intent recognition [9].

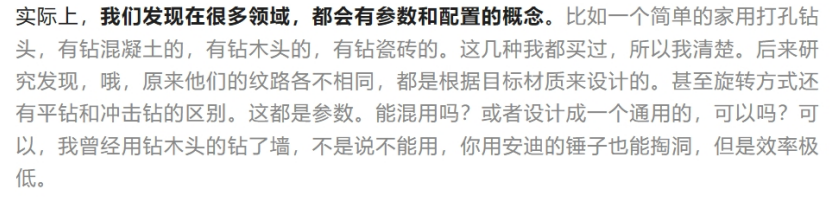

Figure 4: Orignal Picture.

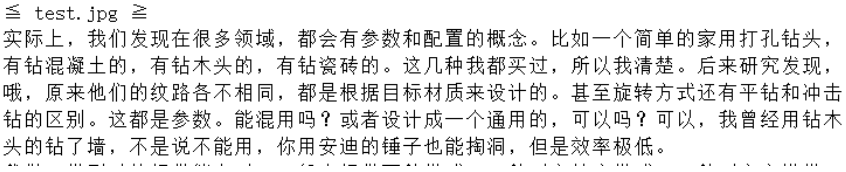

Figure 5: OCR recognition in action.

2. **Image Cropping, Rotation, and Color Adjustments**: These image enhancement techniques simulate various changes in user-uploaded images, such as different shooting angles and lighting conditions. This improves the model's adaptability to diverse visual inputs and enhances its accuracy in image recognition [4].

3. **Text Augmentation Techniques**: We also employ text augmentation methods such as synonym replacement and syntactic reordering. These techniques simulate scenarios where users express the same intent using different sentence structures, thereby improving the model's ability to handle diverse text inputs while maintaining intent accuracy [8].

4.3 Knowledge Distillation and Retrieval-Augmented Generation (RAG)

To further improve model performance, especially in tasks requiring extensive background knowledge, we integrate knowledge distillation and Retrieval-Augmented Generation (RAG) techniques.

1. **Knowledge Distillation**: In knowledge distillation, we transfer knowledge from a large "teacher" model to a smaller "student" model. This approach enables the student model to maintain high accuracy while reducing computational resource consumption. Moreover, knowledge distillation helps improve the model's generalization, particularly in recognizing intents specific to the e-commerce domain [3].

2. **Retrieval-Augmented Generation (RAG)**: The RAG approach combines external knowledge bases with generative models. It allows the model to dynamically retrieve relevant information during dialogue generation. By introducing additional background knowledge during the dialogue process, RAG enhances the model's ability to recognize user intents, particularly in e-commerce scenarios where detailed product descriptions and user queries require external knowledge support [5].

4.4 Model Evaluation and Ablation Experiments

To validate the effectiveness of our model, we conducted a series of ablation experiments to analyze the contributions of each component (such as LoRA fine-tuning, data augmentation techniques, and knowledge integration) to the overall performance. Through rigorous experimental design and parameter tuning, our model outperforms baseline models by 5.35% in intent recognition accuracy, particularly in complex e-commerce dialogue scenarios. These results demonstrate the model's improved ability to accurately recognize users' true intents, showcasing the substantial benefits of integrating large language models in multimodal intent recognition tasks.

5 Results and Analysis

5.1 Experimental Setup and Objectives

In this section, we present the experimental setup and the specific objectives of the experiments conducted to validate the effectiveness of the proposed method. The experiments were designed to explore three key factors:

1. **Introducing OCR Training but Not OCR Inference**: This configuration aims to assess the impact of incorporating OCR-based text extraction during training, without yet applying OCR inference during model prediction.

2. **Introducing Both OCR Training and OCR Inference**: This experimental setup introduces both OCR training and inference, allowing us to explore the full potential of OCR integration, including inference on new, unseen data.

3. **Adjusting LoRA Parameters (Rank=16, Scaling Factor=32) and Introducing OCR Inference**: This setup investigates how optimizing the LoRA (Low-Rank Adaptation) parameters—specifically, adjusting the rank and scaling factor—affects model performance when combined with OCR inference.

For these experiments, the Qwen2-VL-7B model was used, fine-tuned with the Llama-Factory framework for inference. All experiments were performed on A100 and H100 GPUs to ensure consistent computational resources and reproducibility.

	F1	Precision	Recall
Overall	**0.7882**	0.8093	0.787
Dialogue Intent Classification Task	0.8648	0.8812	0.866
Image Scene Classification Task	0.7116	0.7373	0.708

Figure 6: Baseline performance with the initial setup, representing the performance before the introduction of OCR and LoRA optimizations.

5.2 Experiment Results Presentation

In Table 1, we summarize the performance of the model across different configurations. The table presents the results for three primary experimental setups, showing the performance for each epoch in terms of intent recognition, image scene understanding, and the overall average score.

Table 1: Summary of Experimental Results

Configuration	Epoch	Intent Score	Image Scene Score	Average Score
Introducing OCR Training but Not OCR Inference	3.5	86.77	77.47	82.12
	4	88.08	76.82	82.45
	4.5	88.74	77.59	83.17
	5	88.28	77.38	82.83
	5.5	88.07	77.30	82.69
	6	87.87	77.15	82.51
Introducing OCR Training and OCR Inference	4.5	87.84	78.06	82.95
	5	88.28	79.68	83.98
	5.5	88.07	79.04	83.56
	6	87.87	78.98	83.43
Adjusting LoRA Parameters (Rank=16, Scaling Factor=32) and Introducing OCR Inference	5	86.10	79.34	82.72
	5.5	86.28	80.19	83.24
	6	86.94	81.4	84.17
	6.5	85.32	80.28	82.80
	7	85.31	80.86	83.09

5.3 Experiment Results Analysis

5.3.1 Introducing OCR Training but Not OCR Inference

In this configuration, the objective was to evaluate the impact of OCR training on the model's ability to recognize intents from multimodal inputs, without applying OCR inference during the prediction phase.

As shown in Table 1, the highest average score (83.17) was achieved at epoch 4.5, which was slightly better than the other epochs. The intent score reached its peak at 88.74 at epoch 4.5, indicating that OCR training significantly improves the model's intent recognition capabilities.

However, the image scene score remained relatively lower (77.59) in comparison to configurations that involved OCR inference. This suggests that while OCR training improves textual intent recognition, the model still faces challenges in effectively handling image-based content without OCR inference.

Conclusion: The results indicate that OCR training enhances the model's ability to recognize user intent, but without OCR inference, the improvement in image scene understanding remains limited.

5.3.2 Introducing OCR Training and OCR Inference

In this experiment, both OCR training and OCR inference were introduced. This configuration aimed to fully leverage OCR technology to process both the textual and visual elements in the data.

The results in Table 1 show a significant improvement in performance when OCR inference was added. The model reached State-of-the-Art (SOTA) performance at epoch 5, with an average score of 83.98. The intent score was 88.28, and the image scene score improved to 79.68.

This improvement confirms that OCR inference plays a critical role in enhancing the model's multimodal understanding, particularly when it comes to interpreting images. The performance boost is especially noticeable in image scene understanding, where the score increased by approximately 2.3 points compared to configurations without OCR inference.

Conclusion: OCR inference provides a substantial and consistent performance gain, particularly in enhancing image scene understanding, and plays a critical role in the model's ability to process multimodal data effectively.

5.3.3 Adjusting LoRA Parameters (Rank=16, Scaling Factor=32) and Introducing OCR Inference

This configuration involved tuning the LoRA parameters while keeping OCR inference active. The aim was to assess whether optimizing the rank and scaling factor could further improve model performance.

As shown in Table 1, the best performance was achieved at epoch 6, with an average score of 84.17, which is an improvement over the SOTA performance achieved with OCR training and inference alone (83.98). The intent score slightly decreased to 86.94, but the image scene score improved significantly to 81.4, indicating that the LoRA parameter optimization had a notable impact on image scene understanding.

Conclusion: LoRA parameter optimization enhances image scene understanding, and when combined with OCR inference, it significantly improves the overall performance of the model.

5.4 Comprehensive Performance Comparison

In Table 2, we present a comprehensive comparison of the different experimental configurations. This table shows the SOTA score for each configuration, as well as the average score improvement compared to the baseline model.

Table 2: Comprehensive Performance Comparison

Configuration	SOTA Score	Average Score Improvement
Baseline	78.82	-
Introducing OCR Training but Not OCR Inference	83.17	+4.35
Introducing OCR Training and OCR Inference	83.98	+5.16
LoRA Parameter Optimization + OCR Inference	84.17	+5.35

Conclusion Summary:

- **OCR Training and Inference** resulted in the highest performance gains, especially for tasks involving image scene understanding, significantly improving the model's overall performance.
- **LoRA Parameter Optimization** further boosted the image scene score, demonstrating that fine-tuning the rank and scaling factor can maximize model performance.

5.5 Improvement Directions

Based on the experimental results, the following optimization directions are proposed for future work:

- **OCR Optimization**: Enhance OCR text extraction by filtering out irrelevant content and denoising key information to minimize inference interference.
- **Retrieval-Augmented Generation (RAG)**: Incorporate external knowledge bases to enhance model inference, especially in tasks involving detailed product information.
- **Data Augmentation and Parameter Tuning**:
  - Refine datasets with imbalanced labels.
  - Expand the dataset, focusing on augmenting low-scoring labels (e.g., for color coverage or random noise).
- **Chain-of-Thought (CoT)**: Introduce reasoning methods to structure multi-turn dialogues, enhancing the model's ability to reason through complex tasks.

5.6 Summary

Through experimental validation, the proposed method demonstrated substantial performance improvements in multimodal intent recognition:

- After introducing OCR training and inference, the model achieved SOTA with an average score of 83.98.
- With LoRA parameter optimization (rank=16, scaling factor=32), the model achieved the best result of 84.17.

Future work will focus on further optimizing OCR text processing, refining parameters, integrating RAG and Chain-of-Thought techniques, and enhancing the model's generalization and robustness in more complex scenarios.

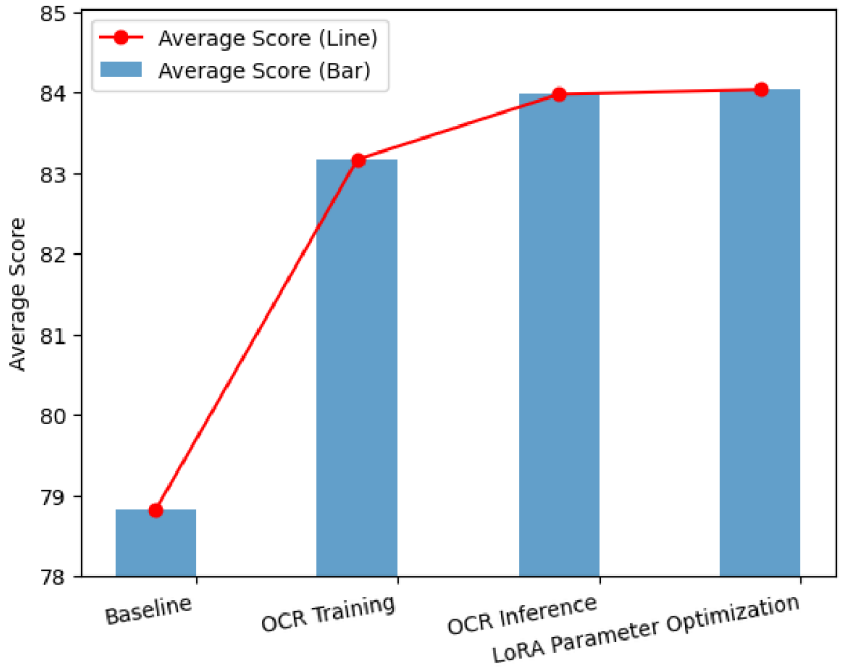

Figure 7: Performance improvement with different configurations, showing the increase from baseline (78.82) to the best result (84.17).

6 Conclusion

This study addresses the challenge of intent recognition in multimodal dialogue systems within the e-commerce domain by proposing an innovative method based on large language models (LLMs).

As user expressions on e-commerce platforms become increasingly diverse, incorporating multiple input modalities such as text, voice, and images, accurately understanding user intent and responding efficiently has become crucial for enhancing user experience and optimizing system performance. However, existing mainstream large language models, despite their excellent multitask learning capabilities in open domains, still exhibit significant shortcomings in domain-specific tasks, particularly in complex e-commerce dialogue scenarios. Issues such as inaccurate intent recognition, severe information omission, and poor context correlation not only degrade user experience but also impact customer satisfaction and sales conversion rates on e-commerce platforms.

To overcome these challenges, this study proposes fine-tuning advanced large language models (e.g., Qwen2-VL) using Low-Rank Adaptation (LoRA) to better adapt to the specific needs of the e-commerce domain. LoRA, as an efficient fine-tuning method, introduces low-rank matrices to adjust model parameters without requiring full retraining, significantly enhancing model performance on domain-specific tasks. Additionally, to address the limitations posed by insufficient data and modality heterogeneity, this study incorporates a variety of data augmentation techniques. These techniques include multiple enhancement strategies for both text and images. For text data, methods such as synonym replacement, syntactic reordering, and random deletion of key characters are used to help the model handle the diversity of user expressions. For image data, techniques like cropping, rotation, color adjustments, and noise addition are employed, along with Optical Character Recognition (OCR) to extract key information from images. These methods effectively mitigate information loss during image-text fusion, expanding the dataset from hundreds to thousands of samples and improving the model's robustness and generalization ability.

Inspired by knowledge distillation and Retrieval-Augmented Generation (RAG) techniques, this study further integrates large language models with external knowledge bases. Knowledge distillation optimizes the model's learning process through a teacher-student framework, allowing lightweight models to retain high performance while reducing computational overhead. RAG technology dynamically retrieves external information (e.g., product descriptions, user reviews, and frequently asked questions) during dialogue generation, providing contextual support for the model's reasoning, which improves both the accuracy of intent recognition and the precision of responses. This multimodal fusion of information, combined with knowledge expansion, enables the model to more effectively capture complex user intents in e-commerce scenarios.

Through rigorous ablation experiments and parameter tuning, this study validates the effectiveness of the proposed method. Specifically, by combining OCR inference with parameter adjustments, the model's average performance improved by 5.35 percentage points, compared to the baseline method. Both intent recognition and image scene understanding scores showed stable gains. These results demonstrate that fine-tuning large language models effectively, while integrating multimodal data and external knowledge, significantly enhances the intent recognition capabilities of multimodal dialogue systems in the e-commerce context.

In conclusion, this study proposes a multimodal intent recognition method tailored to the e-commerce domain, addressing the shortcomings of general models through fine-tuning and data augmentation strategies. Future work will focus on optimizing OCR result filtering mechanisms, exploring the integration of RAG and Chain-of-Thought reasoning, and further improving the intelligence and robustness of multimodal dialogue systems to provide more precise, personalized services and commercial value for e-commerce platforms.

References

[1] H. Chen et al. "A survey on dialogue systems: Recent advances and new frontiers". In: *ACM SIGKDD Explorations Newsletter* 19.2 (2017), pp. 25–35.

[2] E. D. Cubuk et al. "Randaugment: Practical automated data augmentation with a reduced search space". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2020, pp. 702–703.

[3] G. Hinton. "Distilling the Knowledge in a Neural Network". In: *arXiv preprint arXiv:1503.02531* (2015).

[4] A. Koschan and M. Abidi. *Digital color image processing*. John Wiley & Sons, 2008.

[5] P. Lewis et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474.

[6] J. Ni et al. “Recent advances in deep learning based dialogue systems: A systematic survey”. In: *Artificial Intelligence Review* 56.4 (2023), pp. 3055–3155.

[7] D. Ramachandram and G. W. Taylor. “Deep multimodal learning: A survey on recent advances and trends”. In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 96–108.

[8] C. Shorten, T. M. Khoshgoftaar, and B. Furht. “Text data augmentation for deep learning”. In: *Journal of Big Data* 8.1 (2021), p. 101.

[9] A. Singh, K. Bacchuwar, and A. Bhasin. “A survey of OCR applications”. In: *International Journal of Machine Learning and Computing* 2.3 (2012), p. 314.

[10] A. X. Yang et al. “Bayesian Low-Rank Adaptation for Large Language Models”. In: *arXiv preprint arXiv:2308.13111* (2023). URL: https://arxiv.org/abs/2308.13111.