Language Model Alignment for Conversational Shopping at Amazon

Chen Luo
Dimitris Papadimitriou
Hariharan Muralidharan
Amazon
Palo Alto, USA
cheluo@amazon.com

Dhineshkumar Ramasubbu
Aakash Kolekar
Wenju Xu
Cong Xu
Amazon
Palo Alto, USA

Anirudh Srinivasan
Mukesh Jain
Qi He
Amazon
Palo Alto, USA

Abstract

The rapid growth of online shopping stores, such as Amazon, has led to services reaching billions of people worldwide. With global retail sales exceeding $6 trillion in 2024, customer expectations for personalized and seamless shopping experiences have heightened. Traditional online shopping experiences, such as search and navigation systems, often fall short in addressing complex shopping journeys. Conversational shopping (such as Amazon Rufus) offers a transformative approach by enabling dynamic, multi-turn dialogues that closely resemble human interactions. This allows customers to explore product options, seek clarifications, and receive personalized recommendations, thereby enhancing product discovery and informed decision-making. In this paper, we share our year-long journey of using language models for conversational shopping at Amazon and introduce how we use LLM fine-tuning techniques to enhance LLMs for a conversational shopping experience like Amazon Rufus. We also introduce innovative strategies for training data collection and demonstrate real-world applications, including product recommendations, clarification mechanisms, and internationalization for global customers.

CCS Concepts

• **Computing methodologies** → **Natural language generation**; *Machine learning*; • **Information systems** → *Information retrieval*.

Keywords

Large Language Model, Model Alignment, Product Search

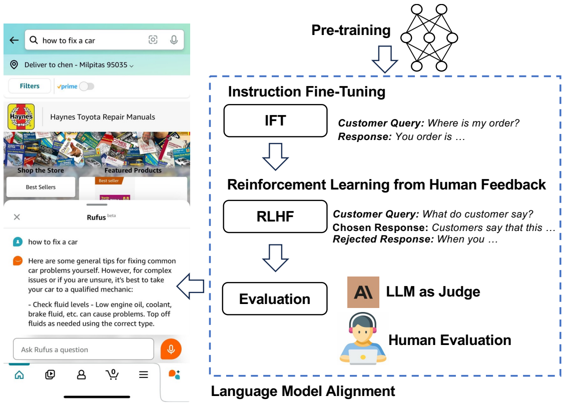
ACM Reference Format:
Chen Luo, Dimitris Papadimitriou, Hariharan Muralidharan, Dhineshkumar Ramasubbu, Aakash Kolekar, Wenju Xu, Cong Xu, and Anirudh Srinivasan, Mukesh Jain, Qi He. 2025. Language Model Alignment for Conversational Shopping at Amazon. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3731955

1 Introduction

The proliferation of online shopping stores, such as Amazon, Walmart, Etsy, and Alibaba, has revolutionized the retail landscape,

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3731955



how to fix a car

Here are some general tips for fixing common car problems yourself. However, for complex issues or if you are unsure, it's best to take your car to a qualified mechanic:

- Check fluid levels - Low engine oil, coolant, brake fluid, etc. can cause problems. Top off fluids as needed using the correct type.

how to fix a car
Deliver to chen - Milpitas 95035
Filters
prime
Haynes Toyota Repair Manuals
Shop the Store
Featured Products
Best Sellers
Best seller
Rufus beta
Ask Rufus a question
Pre-training
Instruction Fine-Tuning
IFT
Customer Query: Where is my order?
Response: You order is …
Reinforcement Learning from Human Feedback
RLHF
Customer Query: What do customer say?
Chosen Response: Customers say that this …
Rejected Response: When you …
Evaluation
LLM as Judge
Human Evaluation
Language Model Alignment

Figure 1: Language Model Alignment for Conversational Shopping at Amazon

providing billions of users worldwide with convenient access to a vast array of products. In 2024, global retail e-commerce sales are projected to surpass 6 trillion, accounting for over 20% of total retail sales [5]. As e-commerce continues to grow, customer expectations for personalized and seamless shopping experiences have also risen. Traditional search and navigation methods, while effective, often fall short in addressing complex shopping journeys where customers explore multiple options, seek recommendations, and require clarifications before finalizing a purchase decision. A recent survey [1] revealed that Americans spend an average of 62 minutes per week shopping online, totaling approximately 54 hours annually, indicating the significant time investment consumers make in online shopping [5] and underscores the need for more interactive shopping solutions to enhance the customer experience.

Conversational shopping such as Amazon Rufus (Fig 1) offers an innovative approach to providing efficient and interactive shopping solutions, transforming the customer experience by enabling more natural and intuitive interactions. Unlike traditional static search queries that require users to precisely articulate their needs in a single attempt, conversational systems facilitate dynamic, multi-turn dialogues that closely mirror human conversations. This allows customers to ask follow-up questions, request clarifications, and explore different product options seamlessly. For example, a customer searching for a laptop can engage in a conversation to specify preferences such as budget, brand, screen size, or intended

[1] https://nypost.com/2024/08/07/lifestyle/americans-spend-more-than-two-days-online-shopping-per-year-study/

use (e.g., gaming, work, or school). The conversational system can then suggest relevant products, highlight key features, and compare alternatives based on the customer's feedback. This interactive process enhances product discovery and empowers customers to make informed purchasing decisions with personalized recommendations. Conversational shopping also reduces friction by addressing customer queries in real time, building confidence and satisfaction. As a result, these systems are becoming a preferred method for online retailers to engage customers and improve conversion rates.

In this paper, we present our year-long effort to use language models [1] for conversational shopping at Amazon. We detail our use of advanced language model alignment techniques, including Instruction Fine-Tuning (IFT) and Direct Preference Optimization (DPO), to enhance LLMs for applications at Amazon Rufus. Our approach includes innovative data collection strategies to support alignment tasks and real-world deployment in product recommendations, clarification questions, and internationalization for global customers. This work highlights the potential of aligned LLMs to make conversational shopping more interactive, personalized, and effective while providing insights into the practical challenges of applying advanced language models in conversational shopping.

2 Model Alignment for Shopping

When building customer-facing conversation shopping applications with LLMs it is crucial to ensure that the generated responses are relevant, accurate, and compliant with trust and safety regulations for effective conversations, as these interactions directly impact customer satisfaction and business outcomes. This underscores the importance of fine-tuning and carefully controlling LLMs before deploying them to customers. To enable the LLM to process product-related information and answer a variety of user questions in different formats, we adopt a two-stage process for LLM alignment in conversational shopping at Amazon Rufus (as in Fig. 1): (1) IFT [10] – Teaching the model to understand product catalogs in the form of evidence. (2) DPO training [7] – Further refining the quality of responses to ensure natural and coherent conversations while delivering high-quality, trustworthy interactions that meet customer expectations.

2.0.1 *Instruction Fine-Tuning.*

The fine-tuning process utilizes carefully structured data pairs consisting of inputs and targeted outputs. The input comprises a prompt containing product or customer information necessary to answer the question, while the output is the desired model response in the correct format. During fine-tuning, the model's performance is optimized by calculating the cross-entropy loss between the model's predicted next-token logits and the actual target tokens [10]. To ensure training stability and enable performance optimization, model checkpoints are saved at regular intervals in our production training stack. These checkpoints serve multiple purposes: selecting the best-performing model for alignment, providing recovery points in case of training interruptions, and enabling performance comparisons across different training stages.

2.0.2 *DPO Fine-tuning.*

IFT adapts LLMs to perform specific tasks by training them on desired outputs. However, this approach does not explicitly prevent the LLM from generating undesirable responses. DPO, an extension of IFT, addresses this limitation by incorporating both positive and negative examples in the training process. The primary objective of DPO is to steer the LLM away from generating responses that fail to meet specified criteria or business objectives. To improve scalability, we use non-fine-tuned LLMs to produce negative responses, such as duplicate product recommendations, abrupt or incomplete answers, and unnecessarily verbose responses. Similar to IFT, checkpoints saved during DPO alignment are evaluated against offline benchmarks, and the best checkpoint is selected for human evaluation and implementation in customer-facing applications.

2.1 Training Data Collection

In each model release, data collection for both IFT and DPO training is determined by any deficiencies identified in specific features in the production model. As deficiencies are identified, feature owners responsible for a particular feature start preparing IFT and DPO datasets for model alignment. Examples of such features include product question answering, product recommendation responses and clarification questions, among others. These newly developed datasets are usually added to the old datasets used in fine-tuning previous model releases

We tailor our techniques for generating IFT and DPO data based on specific features. Model inputs are typically selected from past production data or generated by prompting independent LLMs. For IFT response generation, our editorial teams play a crucial role in providing high-quality responses. These carefully crafted responses align the conversational shopping model with desired behavior. DPO data requires both a "chosen" and a "rejected" response [8]. In some cases, we obtain both responses by prompting an independent LLM to generate them [2, 4]. Another effective approach involves generating an initial, weaker response via a LLM and refining it based on editorial team feedback. The refined version becomes the "chosen" response, while the original serves as the "rejected" one. For example, to adapt the model to regional preferences, chosen and rejected responses may differ in their use of measurement units. To ensure that the sizes of IFT and DPO datasets remain balanced, we design independent fine-tuning experiments that incorporate different proportions of newly generated datasets—typically added to the fine-tuning data from the previous release. These experiments explore a range of dataset compositions, from using none of the newly acquired data to utilizing it fully, along with various subsampling and combination strategies.

2.2 Evaluation

In the development of large language models for conversational shopping, robust evaluation frameworks serve as the cornerstone of ensuring model reliability, safety, and effectiveness. Evaluations are particularly crucial in shopping contexts where model outputs directly influence customer purchasing decisions and satisfaction. While traditional NLP metrics like BLEU [6] or ROUGE [3] are insufficient for capturing the nuanced requirements of shopping conversations, establishing domain-specific evaluation criteria becomes essential for measuring both model performance and business impact.

Table 1: Evaluation of Fine-tuned Models

Metric	Prod	Gamma
System Prompt Accuracy	89.07%	+4.37%
Instruction Prompt Accuracy	69.90%	-1.20%
Consistency Rate	92.08%	+4.23%
STP Compliance Rate	76.35%	+6.08%
Single-Turn Response Format	99.81%	0.00%
Multi-Turn Response Format	98.63%	+4.11%
Multi-Turn Context Carryover	69.49%	-4.23%
Multi-Turn Anaphora Accuracy	81.81%	-3.03%

In our evaluation framework, we established a comprehensive two-tier approach combining offline and online metrics to assess model performance and alignment for conversational shopping. The offline evaluation consisted of Priority Zero (P0) metrics which combined outputs from two parallel tracks: LLM-based judging and human evaluation, both examining various customer-facing features. For conversation quality, we implemented Multi-Turn Context Carryover Coverage and Multi-Turn Anaphora Accuracy Rate to ensure coherent dialogue flows. Notably, we specifically designed metrics to address Trust & Safety with STP Compliance Rate and Single-Turn/Multi-Turn Response Format evaluations serving as key indicators for model usage readiness.

The combination of offline and online evaluation mechanisms not only accelerated our experimental velocity but also ensured that our alignment efforts remained grounded in actual customer needs and behaviors, leading to measurable improvements in helpfulness scores and expert voice ratings in real-world usages. We employ a range of evaluation metrics to assess the performance of our model across different dimensions:

- **System Prompt Accuracy**: Measures adherence to product-related system prompts.
- **Instruction Prompt Accuracy**: Evaluates how well the model follows new instructions.
- **Consistency Rate**: Checks factual alignment between responses and input evidence.
- **STP Compliance Rate**: Ensures responses follow Rufus's trust and safety principles.
- **Single-Turn Response Format**: Verifies inclusion of required CX components in single-turn replies.
- **Multi-Turn Response Format**: Assesses CX component retention in multi-turn interactions.
- **Multi-Turn Context Carryover**: Evaluates the model's ability to retain prior context.

Table 1 shows the evaluation results of the fine-tuned model (Gamma) against the production model across various metrics. We release new models based on these evaluations.

3 Conversation Shopping Experience

In this section, we present three examples of customer experiences to showcase how we empower the fine-tuned language model to enhance real customer experiences at Amazon.

Table 2: Clarifying Questions on Customer Engagement

Metric	Baseline (%)	With CQs (%)	bps+
Click-Through Rate (CTR)	19.2	23.5	+431
Next-Turn Rate	24.9	37.0	+1210

3.1 Empowering Clarifying Questions

Customers often begin their online shopping journey with broad queries that rarely capture their specific intent (e.g., "baking"). Generating Clarifying Questions (CQs) helps resolve these ambiguities by presenting open-ended options tailored to different shopping scenarios. The Clarifying Questions feature bridges this gap by proactively surfacing follow-up questions at the right moment, enabling a deeper understanding of customer needs and preferences. By prompting users for additional details about their broad or generic queries, the LLMs can generate more precise and relevant recommendations in subsequent interactions. This refined approach effectively narrows the customer's product search, enhancing the overall helpfulness of LLM responses.

We concluded the fine-tuning and instruction prompting of the LLM model to empower this feature. We conducted an evaluation experiment (Table 2) to compare the performance of asking clarifying questions against a baseline where no CQs are asked. We observed increased customer engagement with CQs, with the CTR (Total clicks / Total conversations) trending +431 bps higher (23.5% vs 19.2% baseline) and the next-turn rate (Follow-on queries / Total initial queries) trending +1210 bps higher (37% vs 24.9% baseline) compared to the baseline.

3.2 Product Recommendations

Product recommendations are a cornerstone of the conversational shopping experience, bridging the gap between customer intent and product discovery. Unlike traditional search-based shopping, where customers manually refine queries to find relevant products, conversational product recommendation LLMs dynamically interpret user intent, retrieve relevant products, and present tailored suggestions. Our recommendation system is designed to understand customer queries in real-time, decompose them into meaningful aspects, and surface high-quality products that align with user needs. This aspect-driven retrieval ensures that the recommended products are not only relevant but also optimized for discoverability and decision-making.

To enhance accuracy and relevance, we adopt a structured approach that involves three key stages: aspect-based retrieval, product selection, and explanation generation. The first stage, aspect-based retrieval, involves breaking down user queries into multiple dimensions or interpretations. For example, a customer searching for "best headphones" may have different implicit preferences such as "best noise-canceling headphones," "best budget wireless headphones," or "best headphones for workouts." Our system automatically generates these aspects via a LLM and calls the search engine to retrieve a set of products corresponding to each refined search direction. This ensures a diverse and comprehensive set of product recommendations tailored to different user preferences. The second stage, products selection, refines the list of retrieved products, also

Table 3: The product recommendation model comparison

Category	Prod (%)	Gamma (%)
Best Products Rate	34.7%	66.3%
Non-Amazon Products Defects	42%	0.3%
Hallucinations Defects	41%	2%
Missing Best Products	66%	24.8%
Irrelevant Recommendations	35%	2%

via a LLM, by applying structured evaluation criteria. A product is deemed a "decently good" recommendation based on several factors, including relevance to the original query and aspect, brand reputation, customer ratings, and expert endorsements. Well-known brands and top-rated products are prioritized, while lower-quality or lesser-known options are filtered out. This approach prevents the inclusion of misleading or sub-optimal recommendations and instead focuses on delivering trustworthy and high-quality suggestions. Our model also considers whether a product aligns with commonly recognized industry standards—such as Apple AirPods Pro being a default high-quality choice for "best wireless earbuds."

Table 3 presents the performance of our three-stage product recommendation strategy. The updated model (Gamma) generated "Best Products" responses in 66.3% of cases, compared to only 34.7% for the previous production model. The improvements were driven by three key factors: (1) reducing defects related to non-store products and hallucinations (from 42% and 41% to <0.3% and <2%, respectively), (2) decreasing the frequency of "missing best products" from 66% to <25% by refining branded search keyword generation through DPO, and (3) lowering the occurrence of irrelevant product recommendations from 35% to 22% by incorporating additional evidence and improving product selection.

3.3 Multilingual Conversational Shopping

Most online e-commerce stores, like Amazon, serve customers worldwide. Here, we introduce how we adapt our LLM-based conversational shopping experience to accommodate different linguistic, cultural, and behavioral patterns across various countries. The LLM used in our experiments is a model pre-trained on 68 different languages. The LLM already captures some multilingual ability. We additionally performed multilingual alignment fine-tuning and cross-lingual prompting to empower the multilingual conversation shopping experience.

3.3.1 Multilingual Alignment. We conducted two stages of multilingual aware post training 1. Instruction Fine-tuning (IFT) 2. Direct Preference Optimization (DPO) to help the model become more performant in multilingual scenarios. Instead of translating the entire prompt from English to the target language, we proposed to use Language Instructions as a separate component in the prompt where the majority of the language/locale specific instructions would reside. This approach helps to quickly scale, re-use and maintain our suite of prompts across different experimental scenarios. Inspired by the "Pinch of Multilinguality" approach [9], we utilized a small, high-quality multilingual instruction dataset to leverage our strong pre-trained multilingual model's capabilities. We created this limited volume of multilingual IFT data by translating English content and synthesizing new data with language-specific instructions, enabling effective use of the model's multilingual abilities with minimal additional training.

Table 4: Average localization error rates for different localization dimensions.

Model	Spelling	Distance	Date	Speed
w/o Alignment, w/o Prompt	61.9	10.0	6.3	93.8
w/o Alignment, w/ Prompt	36.5	0.0	25.0	0.0
w/ Alignment, w/ Prompt	16.57	10.0	0.0	0.0

3.3.2 Cross-lingual Prompting. Having utilized multilingual data in pre-training and post-training, we started developing the prompts to provide the response in a target language (locale) with the right localization (currency, units of measurement and terminology) for the locale. While we saw that the off-target language rates and localization errors were greatly reduced, the model was still struggling to follow the instructions and we would still get the response in either an incorrect language or with localization errors. To combat this, we reinforced some of the instructions in multiple prompt sections. Until this point, all instructions in the prompt were in English. However, in our experiments, the identity instruction section rarely required tuning so we tried translating it to the target language for reducing off-target language errors. Unexpectedly, we also saw some reductions in hallucinations. We also observed a similar reduction in hallucination rate when using an instruction like "Think in English, respond in target language" right before the question (even though the model generated the response in one shot).

Table 4 demonstrates the effectiveness of alignment and cross-lingual prompting in enhancing the conversational shopping experience. The results show significant performance improvements when applying multilingual fine-tuning and cross-lingual prompting. The "w/o Alignment, w/o Prompt" model lacks multilingual alignment data, whereas the "w/ Alignment, w/ Prompt" model incorporates our technique of adding multilingual language instructions. The tuned prompt includes language instructions utilized during the model's post-training process.

4 Conclusion

In this paper, we outlined our year-long effort to leverage language models for conversational shopping at Amazon. We introduced innovative data collection strategies to support alignment tasks and facilitate real-world deployment across key areas such as product recommendations, clarification questions, and internationalization for global customers. Our work demonstrates the significant potential of aligned LLMs to transform conversational shopping, making it more interactive, personalized, and effective.

Speaker Bio Chen Luo is an Applied Scientist at Amazon Search. He received his PhD from Rice University. His main research interests are in scalable machine learning with application in information retrieval. He publishes at ML and IR conferences and Journals such as WWW, KDD, SIGIR, and JMLR, and regularly serves as PCs for NeurIPS, ICML, KDD and WWW. He has given invited talks at many conferences, workshops and meetups.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[2] Xu Guo and Yiqiang Chen. 2024. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. *arXiv preprint arXiv:2403.04190* (2024).

[3] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.

[4] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126* (2024).

[5] Chen Luo, Xianfeng Tang, Hanqing Lu, Yaochen Xie, Hui Liu, Zhenwei Dai, Limeng Cui, Ashutosh Joshi, Sreyashi Nag, Yang Li, Zhen Li, Rahul Goutam, Jiliang Tang, Haiyang Zhang, and Qi He. 2024. Exploring Query Understanding for Amazon Product Search . In *2024 IEEE International Conference on Big Data (BigData)*. IEEE Computer Society, Los Alamitos, CA, USA, 2343–2348. https://doi.org/10.1109/BigData62323.2024.10826015

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

[7] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2338, 14 pages.

[8] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[9] Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 2304–2317. https://doi.org/10.18653/v1/2024.findings-acl.136

[10] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction Tuning for Large Language Models: A Survey. arXiv:2308.10792 [cs.CL] https://arxiv.org/abs/2308.10792