ARR October 2024 EcomScriptBench Author Response

[Link to the paper (submission version)]
[Link to the forum]

Reviewer 1

What is this paper about and what contributions does it make?

In this paper, the authors define a script planning task that involves the automatic generation of plans necessary to achieve a user's objective and the presentation of products to be purchased at detailed steps. They also construct a large-scale benchmark for evaluating this task using LLMs and human annotations. Additionally, it improves the accuracy of product searches by considering purchase intentions.

Reasons to accept

1. The paper is well-written, and the methodology is logically clear, representing solid work.
2. The purchase intention received a high acceptance rate through expert evaluation, supporting the high quality of the constructed dataset.
3. The constructed dataset is very large-scale.

Reasons to reject

1. I am somewhat concerned that the three defined subtasks are all binary classification problems, and the difficulty of the tasks might be exaggerated by forcing LLMs to solve them in a zero-/few-shot manner. Given that annotation data is available, significantly better performance could likely be achieved by splitting some of it as training data and fine-tuning an encoder model (PTLM) such as RoBERTa.

Questions for the Author(s)

1. Suggestions:

a. The results may behave differently depending on the type and domain of the products, so further insights could be gained by investigating the accuracy for each domain and its causes.

2. Questions:
   a. As mentioned in the limitations, since important elements like User Objectives, Scripts, and Purchase Intentions are all generated by LLMs, it is necessary to evaluate how close these are to the ground truth. However, why are only the purchase intentions evaluated by experts, and not the others?
   b. Are the experts mentioned in sections 4.1 to 4.3 (e.g., L312) different from the expert in 4.4? What exactly are these experts specialized in, and why are they defined as experts?
   c. In section 5.2 (2), it mentions fine-tuning with unlabeled training, but how exactly is this learning conducted?

Soundness: 3.5 Overall Assessment: 3.5

Author Response (Reviewer 1)

Thank you for your detailed review and for recognizing the following strengths of our paper:

- The definition of a novel script planning task that combines user goal achievement with step-wise product purchase planning.
- The construction of a large-scale benchmark dataset with high-quality annotations, validated by a strong expert acceptance rate.
- The introduction of a methodology that improves the accuracy of product searches by incorporating purchase intentions.
- The logical clarity and overall quality of the paper, supported by clear writing and solid methodology.

The following paragraphs address your concerns.

Task Design and Impact of Fine-tuning

Thank you for your question. The reason we did not use annotated data for fine-tuning is the limited size of our annotations (5,000 per task). Splitting this data into training, validation, and testing sets (e.g., 8:1:1) would leave only 500 samples for testing per task, which is insufficient for reliable evaluation.

To address your concern, we performed a random split of the annotations into train:dev:test sets in an 8:1:1 ratio and trained the following models. For RoBERTa and DeBERTa, we used a standard classification objective with cross-entropy loss. For LLAMA, we employed a generative objective, training it to generate a binary prediction token. The models were trained on 4,000 annotated samples, validated on 500 samples, and evaluated on the remaining 500 samples. The results are shown below:

Model	Task 1 Acc	AUC	Ma-F1	Task 2 Acc	AUC	Ma-F1	Task 3 Acc	AUC	Ma-F1
RoBERTa-Large	79.18	79.27	78.86	72.26	72.32	71.74	70.26	70.38	69.83
DeBERTa-v3-Large	81.10	80.76	81.03	74.26	74.56	73.78	72.00	71.93	71.99

| Meta-LLaMa-3-8B | 83.48 | 83.38 | 82.64 | 75.75 | 75.52 | 75.73 | 73.06 | 73.33 | 72.84 |
| Meta-LLaMa-3.1-8B | 85.24 | 85.07 | 84.64 | 76.44 | 76.51 | 75.53 | 74.48 | 74.44 | 74.38 |

From the results, we observe that fine-tuning on annotated data improves model performance, likely due to overfitting to the distribution of human annotations. This improvement even surpasses fine-tuning on unlabeled training data (as explained below). However, there remains significant room for improvement, particularly for the last two tasks.

Moreover, collecting a large amount of high-quality annotated training data is extremely cost-intensive and impractical, making it challenging to train models that generalize effectively to unseen cases. As a result, this paradigm remains infeasible in practice, emphasizing the difficulty of our benchmark.

Performance by Product Categories
Thank you for your question. It is indeed interesting to analyze the models' performance across different categories. Below, we present GPT-4o's performance in the product discrimination task for each major product categories:

Category	Accuracy
"Automotive",	64.58
"Beauty_and_Personal_Care",	63.95
"Cell_Phones_and_Accessories",	82.31
"Clothing_Shoes_and_Jewelry",	78.99
"Electronics",	66.15
"Health_and_Household",	62.08
"Home_and_Kitchen",	65.63
"Grocery_and_Gourmet_Food",	82.49
"Industrial_and_Scientific",	79.51
"Office_Products",	67.42
"Patio_Lawn_and_Garden",	82.84
"Sports_and_Outdoors",	76.68
"Tools_and_Home_Improvement",	65.57
"Toys_and_Games"	84.37

From these statistics, we observe that LLMs perform better in certain categories but fall short in others. While we are unable to pinpoint the exact cause for each category, one possible explanation is that some large categories contain many redundant or poorly-described products. This may hinder the LLM's ability to understand the products accurately without the support of visual images. Further discoveries can be made based on our work.

Questions
> As mentioned in the limitations, since important elements like User Objectives, Scripts, and Purchase Intentions are all generated by LLMs, it is necessary to evaluate how close these are to the ground truth. However, why are only the purchase intentions evaluated by experts, and not the others?

In our data collection process, we asked experts to evaluate intentions, as they are not directly assessed in our task formulations. For user objectives and scripts, we evaluate them directly by incorporating AMT annotations (the first subtask involves verifying the correctness of each script and its steps).

Our annotation results show that 94% of the scripts are plausible, and 61% of the total generated scripts are feasible for e-commerce product associations. We use this 61% of annotated scripts in subsequent annotation tasks. This confirms the high quality of the data collected from GPT-4o-mini.

> Are the experts mentioned in sections 4.1 to 4.3 (e.g., L312) different from the expert in 4.4? What

exactly are these experts specialized in, and why are they defined as experts?

The experts involved are the same group who participated in our prompt design and dataset quality evaluation. In our benchmark curation pipeline, we used GPT-4o-mini as the data generator, with few-shot exemplars guiding the generation process. Expert-curated exemplars were incorporated into the prompts to help the LLM understand the generation objective. Collecting exemplars from multiple experts ensures a balanced and fair prompt design. Subsequently, these experts were tasked with evaluating the quality of crowdsourced labels, as described in Section 4.4.

The experts are NLP scientists with extensive experience in e-commerce NLP. They are well-trained in conducting NLP research and are familiar with the e-commerce domain. Due to anonymity concerns, we cannot disclose any specific information about them. In contrast, the crowdsourced workers from AMT are generally considered to have only a basic understanding of AI, NLP, and related fields. It is not guaranteed that they possess expertise in NLP or e-commerce. Therefore, recruiting experts to verify the annotated labels is critical, as they have a deeper understanding of the tasks and can better assess whether the collected labels align with the task requirements and design.

> In section 5.2 (2), it mentions fine-tuning with unlabeled training, but how exactly is this learning conducted?

Thank you for pointing this out. In EcomScriptBench, we annotated only 5,000 data points per task, leaving a significant portion of the dataset unlabeled. Therefore, we aim to leverage these unlabeled data points as training data to evaluate model performance after fine-tuning. To do so, we treat the unlabeled data generated by LLMs as plausible positive samples and use random negative sampling (randomly selecting another ground truth from a different script/step) to create negative samples.

We then train the LLMs using these unlabeled positive data points and the synthesized negative samples with a generative objective. This involves training the model to generate a binary prediction token indicating whether a data entry is correct, as is typically done when training LLMs. We use the same prompts for training and evaluation to ensure alignment between these phases.

After training, we evaluate the models' performance on the annotated test sets of EcomScriptBench, with the results detailed in our paper.

Model	Task 1 Acc	AUC	Ma-F1	Task 2 Acc	AUC	Ma-F1	Task 3 Acc	AUC	Ma-F1
Meta-LLaMa-3-8B	83.18	83.26	82.57	76.71	76.64	75.89	74.90	74.73	74.40
Meta-LLaMa-3.1-8B	83.86	83.94	83.05	77.70	77.87	77.59	75.88	75.58	75.58
Gemma-2-2B	78.14	78.14	77.99	70.00	69.99	69.68	67.63	67.71	67.50
Gemma-2-9B	82.81	82.92	81.89	78.42	78.05	77.89	74.64	74.59	73.93
Mistral-7B-v0.3	83.82	83.86	83.63	78.94	78.96	78.29	76.67	76.61	75.80

We observe that models trained with our unlabeled data show significant performance improvements compared to zero-shot evaluation. This improvement is likely due to training on a much larger dataset, which enhances their generalization and discriminative abilities to assess the correctness of an e-commerce script.

However, when combining these results with fine-tuning on annotated data, we conclude that fine-tuning alone is insufficient to enable LLMs to effectively solve the task of e-commerce script planning. This highlights the challenging nature of our proposed task and dataset.

We sincerely appreciate your valuable advice and hope that our response will assist you in raising your score. Thank you once again!

Reviewer 2

What is this paper about and what contributions does it make?

This paper presents a new task e-commerce script planning, which involves generating scripts that include relevant product recommendations at specific steps where purchases may be necessary. There is currently a shortage of evaluation datasets that combine script planning with product recommendations. To address this gap, the authors design an evaluation benchmark consisting of three discriminative sub-tasks derived from the planning process. Additionally, they introduce a framework to automatically generate product-enriched scripts for evaluation. The experimental results indicate that (L)LMs encounter challenges in addressing these tasks.

Reasons to accept

1. The authors propose a new task with significant potential. This task enhances the shopping experience by streamlining product searches, ultimately saving time for users while also benefiting businesses.
2. The proposed benchmark is both valuable and well-constructed. It addresses the existing gap in datasets for e-commerce script planning. Furthermore, the test data has been annotated by humans and verified by experts, ensuring its quality.
3. The experiments results reveal the challenges associated with the task and the limitation of LLM in addressing this task.

Reasons to reject:

1. Some aspects of the motivation lack supporting evidence. Line [88-100] describe the semantic gap between the planned steps and the search queries. However, the authors don't provide references or experimental results to substantiate the existence of this gap. It would strengthen the argument if the authors conducted a pilot experiment to make the motivation more convincing.
2. Lack of necessary experimental analysis. (1) the GPT series with COT and SC-COT shows underperformance in the zero-shot setting for task script verification, yet the authors do not explain this discrepancy; (2) three PTLM in the zero-shot setting underperform compared to the random baseline. It is unclear whether this result is reasonable and warrants further discussion.

Comments Suggestions And Typos:

1. The term "their purchase intention" used in Line [114-115] is somewhat unclear. I recommend that the authors provide more detail on what this term means and how they

are obtained. Similarly, the term “intention alignment strategy” in line 125 is introduced without adequate context or explanation, which could lead to confusion.

2. The three sub-tasks are all discriminative tasks. Why do the authors try to train the classification tasks using PTLMs on the proposed benchmark (splitting the annotation data)? The classification results on the PTLMs will affect my evaluation of the benchmark's difficulty.

Soundness: 3 **Overall Assessment: 3**

Author Response (Reviewer 2)

Thank you for your detailed review and for recognizing the following strengths of our paper:
- The proposal of a novel and impactful task, e-commerce script planning, which enhances the shopping experience by streamlining product searches, saving users time, and benefiting businesses.
- The creation of a well-constructed benchmark that fills a critical gap in datasets for e-commerce script planning, with high-quality annotations verified by experts.
- The experimental results, which effectively highlight the challenges of the task and the limitations of current LLMs in addressing these challenges.

The following paragraphs address your concerns one by one.

Evidence to Support our Motivation

Thank you for your question. We indeed conducted a pilot study prior to this research, where we used 200 randomly sampled steps from scripts as search queries to examine the search results of search engines. From this pilot study, we observed that approximately 68% of the steps resulted in a very limited selection of products due to a lack of alignment between product titles/metadata and the steps.

For example, consider the step: “Find a reusable bottle that is easy to clean and suitable for carrying both hot and cold beverages.” When entered as a search query, search engines primarily returned generic listings of reusable bottles without explicitly matching key features such as "easy to clean" or "suitable for both hot and cold beverages." This demonstrates the semantic gap between natural language steps and the structured metadata used by search engines, which hinders effective product retrieval.

To address this, more augmented signals for each product are necessary to better tailor product associations for specific use cases. Intention plays a critical role in bridging this gap and improving the alignment between user needs and product recommendations.

Experiment Analysis

Thank you for your question. Regarding the first question, we have provided a detailed error analysis in Section 5.4 to explain why GPT-4o performs poorly under the CoT setting. GPT-4o suffers from significant issues in product understanding, highlighting the need for further enhancement. It is also important to note that CoT does not always guarantee improved performance compared to zero-shot or few-shot prompting. A widely accepted argument in the community is that CoT excels primarily in tasks involving multi-hop or complex reasoning. However, our task primarily requires product and intention understanding, which cannot be easily improved through CoT or SC-CoT.

For the second question, since PTLMs are relatively outdated and have a limited number of parameters

and training data, their zero-shot performance is highly unpredictable when evaluated on data that are significantly different from their training data. This issue is particularly pronounced in the e-commerce domain, where very few PTLMs have been pre-trained effectively. Thus, we believe the extremely poor performance is reasonable. Similar observations have been reported in related e-commerce tasks by Ding et al., where PTLMs often perform randomly and exhibit biases toward choices with semantics similar to their training data, rather than selecting the correct choice for the task at hand.

Wenxuan Ding, Weiqi Wang, Sze Heng Douglas Kwok, Minghao Liu, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Junxian He, and Yangqiu Song. 2024. IntentionQA: A Benchmark for Evaluating Purchase Intention Comprehension Abilities of Language Models in E-commerce. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2247–2266, Miami, Florida, USA. Association for Computational Linguistics.

Comments and Questions

> The term “their purchase intention” used in Line [114-115] is somewhat unclear. I recommend that the authors provide more detail on what this term means and how they are obtained. Similarly, the term “intention alignment strategy” in line 125 is introduced without adequate context or explanation, which could lead to confusion.

Thank you for pointing this out. By using the term purchase intention, we are referencing prior works by Yu et al., 2023, 2024, where purchase intentions refer to mental states in which agents or humans commit to purchasing specific products. This concept explains what the customer believes or wishes to achieve by purchasing a product, approximating their post-purchase actions to align with the steps outlined in a script.

We follow the previous approach to collect these intentions by distilling them from a powerful LLM. Few-shot exemplars are included in the prompt to better demonstrate intentions to the LLM, enabling the scalable collection of intentions across a wide range of products. Intention alignment strategies are explicitly detailed in Section 4.3, which primarily focuses on identifying the product whose intention best aligns with a step in a script. We encourage readers to consult the main body of the paper for a more comprehensive understanding, given the complicated strategy design. However, we will add a brief summary sentence to enhance clarity. Additionally, we will include references and section pointers in the camera ready version to make this clearer.

Yu, C., Wang, W., Liu, X., Bai, J., Song, Y., Li, Z., ... & Yin, B. (2023, July). FolkScope: Intention Knowledge Graph Construction for E-commerce Commonsense Discovery. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 1173-1191).

Yu, C., Liu, X., Maia, J., Li, Y., Cao, T., Gao, Y., ... & Li, Z. (2024, June). COSMO: A large-scale e-commerce common sense knowledge generation and serving system at Amazon. In Companion of the 2024 International Conference on Management of Data (pp. 148-160).

> The three sub-tasks are all discriminative tasks. Why do the authors try to train the classification tasks using PTLMs on the proposed benchmark (splitting the annotation data)? The classification results on the PTLMs will affect my evaluation of the benchmark's difficulty.

Thank you for your question. The reason we did not use annotated data for fine-tuning is the limited size of our annotations (5,000 per task). Splitting this data into training, validation, and testing sets (e.g., 8:1:1) would leave only 500 samples for testing per task, which is insufficient for reliable evaluation.

To address your concern, we performed a random split of the annotations into train:dev:test sets in an 8:1:1 ratio and trained the following models. For RoBERTa and DeBERTa, we used a standard classification objective with cross-entropy loss. For LLAMA, we employed a generative objective, training it to generate a binary prediction token. The models were trained on 4,000 annotated samples, validated on 500 samples, and evaluated on the remaining 500 samples. The results are shown below:

| Model | Task 1 Acc | AUC | Ma-F1 | Task 2 Acc | AUC | Ma-F1 | Task 3 Acc | AUC | Ma-F1 |

|---|---|---|---|---|---|---|---|---|---|
RoBERTa-Large	79.18	79.27	78.86	72.26	72.32	71.74	70.26	70.38	69.83
DeBERTa-v3-Large	81.10	80.76	81.03	74.26	74.56	73.78	72.00	71.93	71.99
Meta-LLaMa-3-8B	83.48	83.38	82.64	75.75	75.52	75.73	73.06	73.33	72.84
Meta-LLaMa-3.1-8B	85.24	85.07	84.64	76.44	76.51	75.53	74.48	74.44	74.38

From the results, we observe that fine-tuning on annotated data improves model performance, likely due to overfitting to the distribution of human annotations. This improvement even surpasses fine-tuning on unlabeled training data (as reported in our paper). However, there remains significant room for improvement, particularly for the last two tasks.

Moreover, collecting a large amount of high-quality annotated training data is extremely cost-intensive and impractical, making it challenging to train models that generalize effectively to unseen cases. As a result, this paradigm remains infeasible in practice, emphasizing the difficulty of our benchmark.

We sincerely appreciate your valuable advice and hope that our response will assist you in raising your score. Thank you once again!

Thank you for your follow-up question and for giving us the opportunity to clarify our statements further.

First, we acknowledge that human annotations are treated as the gold standard for these tasks, and models aligning with human annotations are desirable as they signify high task-specific performance. However, when we mentioned “overfitting to human annotations,” we were referring to the limited diversity in the annotated training set, which consists of only 4,000 samples per task. Compared to millions of unlabeled training examples used in our unlabeled pre-training, this smaller dataset represents a narrower distribution of examples and is more likely to result in overfitting to a limited number of training data. Moreover, while fine-tuning on annotated data improves performance, it is unclear whether this improvement reflects true generalization or simply better alignment with the specific patterns present in the annotations, given the limited scale of the data.

This limited distribution poses two main challenges:

- Generalization to Unseen Data: Fine-tuning on a small dataset with a limited range of cases can lead models to capture task-specific patterns that are disproportionately represented in the training set but may not generalize well to broader, unseen scenarios. While the models perform well on the held-out test set (500 samples from the same distribution), this improvement might not fully reflect their ability to handle more diverse, real-world examples that extend beyond the scope of the annotated data.
- Benchmark Difficulty: One of the goals of our benchmark is to assess whether models can effectively learn task-specific behaviors from broader contexts (e.g., large-scale unsupervised or pseudo-labeled data). Although models fine-tuned on human annotations show improved performance, this improvement highlights the challenge of generalizing beyond limited, expensive-to-create datasets. In practical applications, where such annotated data is often unavailable, models need to demonstrate robust performance without heavy reliance on annotated training sets.

Regarding the phrasing of "even surpasses fine-tuning on unlabeled training data," we acknowledge that it could be confusing without further context. What we intended to emphasize was the relative strength of human-annotated fine-tuning, despite its limited size. This reflects the high-quality nature of the annotations but also underscores the challenges of relying solely on such datasets. While the performance on annotated data fine-tuning is better, the gap to optimal performance (as evidenced by errors and performance variability across tasks) reveals the inherent difficulty of the benchmark.

Finally, we recognize that annotating additional data may improve performance, but the cost and scalability challenges make this an impractical solution for most real-world applications. The gap between the observed performance and real-world expectations demonstrates the benchmark's ability to

stress-test models in realistic, resource-constrained conditions. We will explicitly discuss these in our camera ready version.

Reviewer 3

What is this paper about and what contributions does it make?

The paper proposes the task of e-commerce script planning (ECOMSCRIPT) and introduces a framework for collecting product-enriched scripts. The authors have conducted experiments with various LLMs and advanced prompting methods. The experiments demonstrate the challenges of the task and potential solutions to improve the performance of LLMs on ECOMSCRIPT.

Reasons to accept

3. The experiments with various LLMs and prompting methods show fair results and challenges of the task on ECOMSCRIPT.
4. The authors constructed the large amount of the data (ECOMSCRIPT), which will be available for other researchers.

Reasons to reject

1. Explanation of evaluation metrics is not enough, readers may need more detail.

Questions for the Author(s)

1. The paper is very interesting for the reviewer. This framework can work as knowledge for e-commerce concierges.
2. The reviewer is wondering how the matching works if there is a gap between intention and step, because those two are independently acquired. It could affect the final performance.
3. The authors used Amazon reviews (review sentences + products) in the paper. However, in E-commerce, there are lots of products without reviews or with very poor reviews. In that case, how does this approach work?

Soundness: 4 **Excitement: 3.5**

Author Response (Reviewer 3)

Thank you for your detailed review and for recognizing the following strengths of our paper:
- The introduction of the novel task of e-commerce script planning (EcomScript) and a framework for collecting product-enriched scripts.
- Comprehensive experiments with various LLMs and advanced prompting methods, highlighting both the challenges of EcomScript and potential solutions for improving model performance.
- The creation of a large-scale EcomScriptBench dataset, which will be made available to support further research in this area.

The following paragraphs address your concerns one by one.

 # Explanation of Evaluation Metrics
Thank you for your question. In this work, we design the three subtasks in EcomScript as binary classification tasks to facilitate automated and convenient evaluation. Consequently, we employ three distinct evaluation metrics to ensure that the results for each subtask are comprehensive and unbiased. Specifically, we use accuracy, area under the curve (AUC), and macro F1 score as our metrics.

- Accuracy measures the proportion of correctly classified instances, offering a straightforward evaluation of the model's performance.
- To address class imbalance, we include AUC, which evaluates the model's ability to distinguish between classes.
- Finally, the macro F1 score provides an equal weighting of precision and recall across all classes, ensuring a balanced performance assessment.

These three metrics—accuracy, AUC, and macro F1—jointly capture overall correctness, discriminatory power, and balanced performance, offering a comprehensive evaluation of each subtask in EcomScript.

For each metric, we compute it by comparing the predicted labels from the language models against human-annotated labels, similar to the evaluation process for binary classifiers in traditional machine learning settings. Detailed implementation guidelines are available at: https://scikit-learn.org/1.5/modules/model_evaluation.html#classification-metrics.

 # Comments
> The paper is very interesting for the reviewer. This framework can work as knowledge for e-commerce concierges.

Thank you for your positive feedback! We're delighted to hear that you find our framework interesting and relevant to e-commerce concierge applications.

In general, yes. We believe that the EcomScriptBench not only enhances script planning but also provides valuable insights into product associations based on user intentions. This can significantly improve the shopping experience by enabling more personalized recommendations. Moreover, it paves the way for developing personalized e-commerce shopping assistant agents with robust memory and advanced e-commerce planning capabilities, revolutionizing the customer shopping experience.

> The reviewer is wondering how the matching works if there is a gap between intention and step, because those two are independently acquired. It could affect the final performance.

As explained in our paper, leveraging purchase intention is our proposed solution to approximate the semantic search space between script steps and queries linking to specific products. Using intention as the connecting link is inherently more effective than traditional search queries, such as keywords in product titles and metadata, which are often not represented in script steps.

To best align intentions with steps, we create exemplars with similar semantics and grammatical structures to effectively guide GPT-4o-mini in generating steps and intentions with consistent linguistic patterns (e.g., omitting the subject, using the simple present tense, and keeping them short and concise). However, gaps between intentions and steps can still occur. To address this, we generate 10 intentions per product to ensure as much coverage as possible. In industrial applications, even more intentions per product could be generated to enhance coverage and improve alignment further, given the low cost of generating outputs with GPT-4o-mini.

In our current dataset construction pipeline, expert evaluations and human annotations confirm that our method is effective and does not significantly impact final performance. However, verifying its efficacy at an industrial scale is left to future work by the e-commerce community.

> The authors used Amazon reviews (review sentences + products) in the paper. However, in E-commerce, there are lots of products without reviews or with very poor reviews. In that case, how does this approach work?

This is an excellent question. First, we would like to emphasize that user reviews in our framework are used primarily by the LLM to infer potential user objectives. For example, from a detailed user review of a Nike running shoe, objectives such as participating in a marathon can be inferred.

In cases where reviews are unavailable after quality filtering, we observe that the LLM can still infer these objectives from the product title and metadata alone, even without user reviews, as title and metadata already include sufficient information to reason potential use cases of the product. Therefore, it is theoretically 100% feasible to run our framework without relying on user reviews. However, in our paper, we aim to simulate real-world scenarios as closely as possible. For this reason, we choose to incorporate user reviews to better justify the practical applicability of our framework.

We sincerely appreciate your valuable advice and hope that our response will assist you in raising your score. Thank you once again!

Reviewer 4

What is this paper about and what contributions does it make?

The goal of the work is to facilitate scalable generation of product-enriched scripts by associating products with each step of a script based on the semantic similarity between script actions and purchase intentions. Towards that end, the submission proposes defining the task of generating product-enriched scripts in terms of three sequential subtasks. The hope is that an LLM can perform the three tasks to build a generate-then-discriminate paradigm. A key idea of the submission is to bridge the gap of semantic discrepancies between the planned steps and the search queries intended for search engines , by searching for products using purchase intentions and filtering for products whose purchase intentions align with the planned action. The authors provide product-enriched scripts as an evaluation benchmark.

Reasons to accept

1. The writeup overall is quite easy to follow.
2. The paper uses quite an extensive set of PTLMs and LLms for evaluation.
3. The authors show effectiveness of intention injection.

Reasons to reject

1. Some more detail on the distribution of product categories and its potential impact on results would be useful.

Questions for the Author(s)

- A general question I have concerns the relation of objectives and intentions: how many cases do you have in the data, where an intention more or less matches an objective? If you eliminated such cases from the intentions how would that affect the retrieval of products? Or conversely, how much do non-script-level intentions contribute?

lines 1375-76 Objective: Participate in a maratahon lines 1442-43: Intention 2: Train for a marathon orproduct. There are also steps that can other long-distance race

- lines 356 following: I was wondering what difference it might make to explicitly model the intentions either as forward-looking ("in order to X")= purposes or as backwards-looking "because (in the past) Y" reasons. From your discussion of prompting I have the impression that you are aiming mostly for forward looking purposes as intentions. So I was wondering what would happen if you instead or separately also prompted for backward looking reasons.

E.g. in your example on p. 16 , there are many cases that I could frame either way:

- Intention 3 Increase comfort during daily jogging sessions (forward looking) I've felt discomfort during daily jogging sessions with my old shoes. (Backward looking)
- Intention 4: Reduce the risk of injuries by using shoes with advanced technology (forward looking) (Because) I've had problems with my foot when running in low tech/simple shoes (backward looking)

"We emphasize modeling the customer’s mental state, using phrases like “PersonX wants to buy this because” or “PersonX believes buying this can” to guide the generation."

- lines 058 following From the prompt in Fig 1, I can't see any expectation for eight steps.

"LLM assistant is expected to generate an eight-step script toward the user’s objective, plan an autumn themed party with friends and family"

- You should introduce the name FolkScope when you discuss Yu et al 2023.

" transformed FolkScope into an evaluation benchmark" (line 186)

- Where does this initial script come from? Remind the reader that it's LLM generated.

"Initially, the model is given a user objective o, a script consisting of k steps toward this objective So = {s1 , s2, ..., sk }, and a pool of n e-commerce products P = {p1, p2, ..., pn }." (208-211)

- § 4.1. User Objective and Script Collection: How diverse are the objetives that you collect? For instance, to take the example from Fig 2, does the dataset also include objectives such as "Participate in a half marathon" , "Run at 15k" and more that one might consider to be pretty close.
- What are the details of this filtering? How many words is too short , how many hashtags is too many? Did you experiment with these values?

"To ensure high quality, we discard reviews that are too short or contain excessive punctuation or hashtags." (lines 316-17)

- I recognize that for businesses it may make sense to maximize here, but one might consider this as somewhat ethically problematic: doesn't this maximize consumerism?

"Specifically, we require the LLM to avoid generating overly simple actions and to maximize the necessity of purchases in each step by generating actions that may require items to complete " (327 following)

- I was wondering to what extent the authors' approach could deal with unusual objectives and/or with creative uses of products for purposes that they're not primarily made for. I would expect that on the current setup the products that are associated with user objectives usually are very standard.

"This might be silly to some, but I actually use an automobile windshield shade to cover a very large window in my home! "

- How often did truncation occur?

"For simplicity, we ask the LLM to generate scripts that contain no more than 10 steps, and longer scripts will be truncated to a maximum of 10 steps." (lines 335 following)

- §4.4 Human Annotations: How was IAA for each of the distinct sub-tasks? Does any of the subtasks have significantly lower IAA than the others? Do IAA gradations correlate with system performance?
- §5.4. Some concrete examples of errors would be good to show.
- Re Purchase Intention Mining: considering the prompting in A.1.2, I am wondering how much differentiation in responses you actually get for specific products. I.e. a good deal of the intentions you got for "Nike Air Zoom Running Shoes" seems like it would apply to most other running shoes. Could you not collect general intentions for broader categories (e.g. "Running shoes") and then just add specific goals that only the specific products support?
- Related to previous: I am aware that you use only a 10% subset of the Amazon Review Dataset. But might the dataset not still be skewed with some product categories much larger than others, with corresponding differences in performance per category?

"To manage the over- whelming size of the product pool P and reduce product redundancy, we randomly sample 10% of products from each category while maintaining the original distribution of products"

- A.1.1 Under script collection you show the below on lines 1402-1408. The same also appears as part of a prompt during Step-Intention Alignment (1586 following): is that an error or was there no checking before step intention alignment whehter the script is sound? Did "bake the cake" really pass through script verification?

Step 9: Rest and recover (Take rest days seriously to prevent injury and allow your body to recover)

Step 10: Bake the cake (Place pans in the oven and bake for the time specified in the recipe, checking doneness with a toothpick)

Soundness: 4 **Excitement: 4**

Author Response (Reviewer 4)

Thank you for your detailed review and for recognizing the following strengths of our paper:
- The clear and accessible presentation of our proposed framework for generating product-enriched scripts based on semantic alignment between script actions and purchase intentions.
- The introduction of a novel generate-then-discriminate paradigm using three sequential subtasks to address the challenges of product-enriched script planning.
- The extensive evaluation conducted using a diverse set of PTLMs and LLMs, demonstrating the effectiveness of the proposed approach.
- The validation of intention injection as an effective technique for bridging semantic discrepancies between planned actions and search queries.

The following paragraphs address your concerns one by one.

Performance by Product Categories

Thank you for your question. It is indeed interesting to analyze the models' performance across different categories. Below, we present GPT-4o's performance in the product discrimination task for each major product categories:

Category	Accuracy
"Automotive",	64.58
"Beauty_and_Personal_Care",	63.95
"Cell_Phones_and_Accessories",	82.31
"Clothing_Shoes_and_Jewelry",	78.99
"Electronics",	66.15
"Health_and_Household",	62.08
"Home_and_Kitchen",	65.63
"Grocery_and_Gourmet_Food",	82.49
"Industrial_and_Scientific",	79.51
"Office_Products",	67.42
"Patio_Lawn_and_Garden",	82.84
"Sports_and_Outdoors",	76.68
"Tools_and_Home_Improvement",	65.57
"Toys_and_Games"	84.37

From these statistics, we observe that LLMs perform better in certain categories but fall short in others. While we are unable to pinpoint the exact cause for each category, one possible explanation is that some large categories contain numerous redundant or poorly described products. This may hinder the LLM's ability to accurately understand these products without the support of visual images. Future work could focus on improving LLMs' precise product understanding in weaker categories and enhancing their capabilities as script planners. Additionally, we will include a statistical plot showing the number of products in each category.

Comments

Before we start to respond, we sincerely thank you for these very detailed comments on our paper. We respond to each of them below.

> A general question I have concerns the relation of objectives and intentions: how many cases do you have in the data, where an intention more or less matches an objective? If you eliminated such cases from the intentions how would that affect the retrieval of products? Or conversely, how much do non-script-level intentions contribute?

From the distribution of embedding similarity, we observe that 13% of intentions have a similarity score higher than 0.5 with the user objective. If we were to eliminate cases where intentions closely match objectives, we would likely see a decrease in the effectiveness of product retrieval. This is because many product recommendations depend on the semantic alignment between intentions and the specific steps users need to take within their scripts. Removing these closely matched cases would create gaps in relevant product associations, potentially resulting in less accurate or less relevant recommendations. Therefore, we believe that non-script-level intentions—those that do not directly correspond to a specific step in a script—also play a significant role in enhancing product retrieval and user experience.

> lines 1375-76 Objective: Participate in a marathon lines 1442-43: Intention 2: Train for a marathon of product. There are also steps that can other long-distance race

Yes, there are indeed an infinite number of possible scripts that can satisfy a specific user objective. In our work, we do not aim to exhaust these possibilities but rather to collect a subset of scripts to demonstrate the feasibility and effectiveness of curating a script planning benchmark with e-commerce products. To generate more scripts for a single objective, future work could leverage LLMs to modify each

step into alternative possibilities, thereby increasing the diversity of the scripts.

On the other hand, semantic embedding similarity can address the challenge of matching between related instances, such as marathons and other types of long-distance races. Since these instances have similar embeddings, our framework can still select similar products, ensuring consistency and relevance in product recommendations.

> lines 356 following: I was wondering what difference it might make to explicitly model the intentions either as forward-looking ("in order to X")= purposes or as backwards-looking "because (in the past) Y" reasons. From your discussion of prompting I have the impression that you are aiming mostly for forward looking purposes as intentions. So I was wondering what would happen if you instead or separately also prompted for backward looking reasons.

Thank you for your insightful question regarding the modeling of intentions in our framework. We agree that differentiating between forward-looking purposes and backward-looking reasons is a compelling approach, and it could enrich the understanding of customer motivations.

Our design primarily emphasizes forward-looking intentions—phrased as "in order to X"—to align with the proactive nature of e-commerce interactions. This forward focus mirrors the requirements of most e-commerce contexts, where customers are usually looking to achieve specific goals or outcomes through their purchases. By modeling intentions this way, we aim to create actionable scripts that guide users towards fulfilling their shopping needs effectively. Moreover, our methodology is informed by previous works in the field of e-commerce intentions (FolkScope, COSMO, MIND, IntentionQA), which have largely adopted a similar forward-looking perspective. This alignment ensures that our approach remains consistent with previous works.

That said, we acknowledge the potential benefits of integrating backward-looking reasons into our framework. Capturing past experiences, such as discomfort with previous products, could provide additional context that enhances the relevance of product recommendations. Exploring this dual approach in future iterations of our work could offer valuable insights into consumer behavior, and we appreciate your suggestion to consider it.

> lines 058 following From the prompt in Fig 1, I can't see any expectation for eight steps.

For better visual clarity, we truncated the prompt in the figure to highlight only the key part for demonstration purposes. In the actual prompt, we instruct the LLM to generate a script containing 5-10 steps and allow it to determine the exact number of steps. As a result, the eight-step script shown here is reasonable and deemed plausible by the LLM itself.

> You should introduce the name FolkScope when you discuss Yu et al 2023.

Thank you for pointing this out. This is a typo, and we will ensure the missing reference is included in the camera-ready version.

> Where does this initial script come from? Remind the reader that it's LLM generated.

Yes, we are defining our task by introducing the input and the expected output. The input script is generated by the LLM, as described in our benchmark curation section. We will clarify this in the revised version.

> 4.1. User Objective and Script Collection: How diverse are the objectives that you collect? For instance, to take the example from Fig 2, does the dataset also include objectives such as "Participate in a half marathon" , "Run at 15k" and more that one might consider to be pretty close.

We acknowledge that there are an infinite number of possible, and possibly similar, user objectives for each product. To ensure broader coverage across various domains and product categories, we collect only one objective per product. We believe this approach maximizes the diversity of the collected objectives. However, it is also feasible to collect more than one objective per product, though this would likely result in objectives that are closely related.

> What are the details of this filtering? How many words is too short , how many hashtags is too many? Did you experiment with these values?

We drop reviews that are less than 10 tokens or contain fewer than 3 unique tokens. Additionally, we exclude reviews with more than 5 hashtags, as these are sometimes misleading. These thresholds were determined based on our prior experience in processing e-commerce reviews, and they have proven to provide the best trade-off in retaining the maximum number of valid reviews.

> I recognize that for businesses it may make sense to maximize here, but one might consider this as somewhat ethically problematic: doesn't this maximize consumerism?

Our design is based on the observation that without this constraint, all generated steps tend to be simple and easy to execute, with only around 10% of steps requiring product purchases. This would make our research question difficult to study. To address this, we slightly modified the prompt to better align with our goal of having some steps in the script assisted by e-commerce products. This adjustment resulted in approximately 50% of steps requiring product assistance. It is important to emphasize that we do not aim to maximize consumerism in this manner; our goal is to enable the LLM to be as helpful as possible, functioning as an agent with planning and product recommendation capabilities.

> I was wondering to what extent the authors' approach could deal with unusual objectives and/or with creative uses of products for purposes that they're not primarily made for. I would expect that on the current setup the products that are associated with user objectives usually are very standard.

This is a tricky question. While we do not explicitly ask the LLM to generate creative intentions or consider rare scenarios, such abilities may be limited in the current setup. In our present framework, most products are viewed as functional across a wide range of use cases, as reflected in the collected intentions. To address more unusual or creative use cases, we might need to explicitly instruct the LLM to generate such intentions, which we believe it can achieve with appropriate prompt exemplars and demonstrations.

> How often did truncation occur?

We recorded only 3,098 cases where truncation occurs, which is very rare. In most cases, LLM follows our instruction precisely.

> §4.4 Human Annotations: How was IAA for each of the distinct sub-tasks? Does any of the subtasks have significantly lower IAA than the others? Do IAA gradations correlate with system performance?

The IAA and Fleiss Kappa scores for the three subtasks are closely aligned, with a difference range of ±0.05. We implement strict quality control measures for worker selection and annotation verification, ensuring that all subtasks have sufficiently high IAA. We believe that a high IAA guarantees fair and accurate evaluation, which should not directly impact system performance but rather reflects the model's capabilities.

> §5.4. Some concrete examples of errors would be good to show.

Thank you for your suggestion. We will include an error analysis in the appendix in the camera-ready version.

> Re Purchase Intention Mining: considering the prompting in A.1.2, I am wondering how much differentiation in responses you actually get for specific products. I.e. a good deal of the intentions you got for "Nike Air Zoom Running Shoes" seems like it would apply to most other running shoes. Could you not collect general intentions for broader categories (e.g. "Running shoes") and then just add specific goals that only the specific products support?

When collecting intentions, we provide the LLM with the product title and metadata and ask it to generate intentions based on the product's features. As a result, there are slight differences in intentions between different products. For example, for a Nike running shoe versus a general category of running shoes, intentions like "add to Nike fan's collection" can be incorporated. While it is theoretically feasible to collect broad intentions and then add feature details, this approach has not been verified. Instead, we are learning from the success of previous works, such as FolkScope and COSMO by Yu et al.

> Related to previous: I am aware that you use only a 10% subset of the Amazon Review Dataset. But might the dataset not still be skewed with some product categories much larger than others, with corresponding differences in performance per category?

The dataset is indeed unbalanced across different categories, which is to be expected, as some categories naturally contain more products than others. These broader categories offer more products and contribute to more plans, but they also present additional challenges to the LLM. Specifically, the LLM must have a broader yet precise understanding and differentiation between similar products within the same category. From the table at the beginning, we can infer that category size likely has some impact, although quantifying this effect is not straightforward.

> A.1.1 Under script collection you show the below on lines 1402-1408. The same also appears as part of a prompt during Step-Intention Alignment (1586 following): is that an error or was there no checking before step intention alignment whether the script is sound? Did "bake the cake" really pass through script verification?

When collecting data from GPT, we perform a check between these two stages to ensure that only plausible scripts are further annotated.

We apologize for the typo in the presented script, as it seems we mixed up two scripts. :(We will fix this in our future camera-ready version. Many thanks for pointing this out.

We sincerely appreciate your valuable advice and hope that our response will assist you in raising your score. Thank you once again!

Confidential Comment

Meta Response

Metareview:
This paper introduces EcomScriptBench, a novel multi-task benchmark for e-commerce script planning, which involves generating coherent sequences of actions (scripts) to achieve specific user objectives while associating relevant products at each step. The proposed task, called E-commerce Script Planning (EcomScript), addresses challenges such as:

1. The inability of language models (LLMs) to simultaneously plan scripts and retrieve products.
2. Semantic gaps between planned actions and product search queries.
3. The lack of evaluation benchmarks for this task.

To tackle these issues, the authors define EcomScript as three sequential subtasks and propose a framework that incorporates purchase intentions to enhance product recommendations. The framework uses semantic similarity between actions and purchase intentions to associate products with script steps. The authors also construct EcomScriptBench, the first large-scale dataset for this task, consisting of 605,229 scripts generated from 2.4 million products. A subset of the dataset is annotated by humans and verified by experts to serve as a benchmark. Extensive experiments reveal that existing LLMs, even after fine-tuning, struggle with this task, highlighting its difficulty and the need for further research. The paper has implications for improving e-commerce assistants by enabling better script planning and personalized product recommendations.

Summary Of Reasons To Publish:

1. All reviewers agree that this paper has made qualified contribution.
2. Novel Task Definition: a new and impactful task that combines goal-oriented script generation with product recommendation
3. High-Quality Benchmark Dataset
4. Comprehensive Experiments
5. Clear Writing and Structure

Summary Of Suggested Revisions:

1. Provide Evidence for the Semantic Gap: The paper discusses a semantic gap between planned actions and product retrieval but does not provide experimental or literature-based evidence to support this claim. Adding experiments or references could strengthen the argument.
2. Improve Experimental Analysis
3. Address Task Design Limitations: All subtasks are designed as binary classification problems, which might underestimate model capabilities. Including experiments where models are fine-tuned on labeled data could better evaluate task difficulty.
4. Analyze Product Category Distribution: The dataset may have an imbalanced distribution of product categories, which could affect model performance. A detailed analysis of category-level performance and its impact on results is recommended.

Overall Assessment: 4 = There are minor points that may be revised

We sincerely appreciate the area chair and all reviewers for their recognition of our work, highlighting its qualified contributions in the following aspects:

- The introduction of E-commerce Script Planning (EcomScript) as a unique and impactful task that integrates goal-oriented script generation with product recommendation, addressing key challenges in planning and retrieval.
- The construction of EcomScriptBench, the first large-scale dataset for this task, consisting of over 600,000 scripts derived from millions of products, with a subset annotated and verified by human experts to serve as a reliable benchmark.
- Extensive empirical studies demonstrating the difficulty of the task and revealing that even fine-tuned language models struggle, highlighting the need for further advancements in this domain.
- The clarity and organization of our paper, which effectively conveys the motivation, methodology, and findings of our work.

The area chair mentioned several minor issues that we have duly addressed in our author response, leading to reviewers improving their scores after acknowledging our effective response. Specifically:

- A pilot study using 200 script steps as search queries showed that 68% failed to retrieve relevant products due to misalignment between natural language and product metadata, highlighting the need for augmented signals such as purchase intentions.
- We provided an error analysis explaining GPT-4o's struggles with product understanding and why CoT reasoning does not always help.
- We fine-tuned models on our limited annotated data, observing performance gains but also overfitting, reinforcing the challenge of obtaining large-scale, high-quality annotations.
- We examined model performance across product categories and found significant variations, likely due to redundant or poorly described listings. We will add a statistical plot to illustrate the category distribution.

All these minor issues will be addressed in our camera-ready version.