Supplementary Material

Anonymous submission

Datasets

We conducted experiments through two CRS datasets, DuRecDial2.0 (Liu et al. 2021) and KoRecDial. DuRecDial2.0 is a public English CRS dataset containing 8,241 dialogs and 127,673 utterances related to various domains (*i.e.*, movie, food, POI, and music). KoRecDial is a private non-English CRS dataset collected for researching movie recommendation services, containing 1,000 dialogs and 12,000 utterances. Every system's utterance (*i.e.*, response) in both datasets has a type such as QA, recommendation, and chit-chat. Here, we only use the dialogs whose corresponding responses are of the QA and recommendation types, which should be generated in a factually correct manner. We split the dialogs in both datasets into training, validation, and test sets by the 7:1:2 ratio.

Both datasets employed human workers to annotate each response with related knowledge, indicating relevant factual information associated with the response (Liu et al. 2021). Following (Zhang et al. 2021), we regard this related knowledge as the ground-truth relevant passages for evaluation. Note that these relevant passages contain useful information about the recommended items when the type of response is a recommendation. To construct the passage corpus, we gather all relevant passages in each dataset. Here, the passage corpus for training is constructed solely from the training data. As such, the training passage corpus for DuRecDial2.0 (resp. KoRecDial) contains 8,606 (resp. 3,335) passages. For inference, the DuRecDial2.0 (resp. KoRecDial) passage corpus contains 10,752 (resp. 3,973) passages.

Meanwhile, we could not conduct experiments using *other CRS datasets* (*e.g.*, ReDial (Li et al. 2018) and Inspired (Hayati et al. 2020)) as they do not contain ground-truth relevant passages annotated by humans for evaluating the passage retrieval task.

Implementation

For the *adaptive item selection*, we set the threshold $\sigma_{conf} = 70\%$ during inference time. However, we limit the maximum number of selected items to 3 to prevent excessive noise. Furthermore, to train the passage retrieval module to be robust against noisy items, we intentionally increase the number of selected items by setting the threshold $\sigma_{conf} = 100\%$ during training time.

Input Prompt

Pretend you are a conversational recommender system. Your task is to select the relevant passages for generating a response to the given dialog.

Here is the dialog: {dialog placeholder}

Here is the response: {response placeholder}

Here are the candidate passages: {candidate passages placeholder}

Select the relevant passages from the above passages list that are relevant for generating a response to the provided dialog.
Only answer the relevant passages in order of relevance and do not add an explanation.

Figure 1: Input prompt for passage labeling through GPT-4o.

For the passage retrieval module, we set the number of retrieved relevant passages to $K = 3$, the total number of pseudo-relevant passages to $M = 2$, the dimensionality to $d = 768$, and the batch size to $B = 32$. The maximum length for an enhanced dialog and a passage was set to 128. We also set the learning rate to 1e-5 for DuRecDial2.0 and 1e-4 for KoRecDial. To obtain a hard negative passage, we selected one from the top-20 passages with the highest pseudo-relevance scores, excluding the pseudo-relevant passages Q for the sample. We also obtained the in-batch negative passages by using the hard negative passages from other samples within the same batch.

For the passage labeling, the input prompt used through GPT-4o to rerank the candidate passages is described in Figure 1. For model training, we adopted the Adam optimizer and carefully tuned hyperparameters using the validation set. We implemented our model with PyTorch and conducted experiments on a machine with an NVIDIA A100 GPU.

For the response generation module, the maximum length of input and the batch size B were set to 256 and 4, respectively. We set the learning rate to 1e-5 (for DuRecDial2.0) and 1e-4 (for KoRecDial). We use BART-large (Lewis et al. 2020a) and LLaMa2-7b-chat (Touvron et al. 2023) as generative language models.

The codebase of ESPRESSO will be provided upon acceptance and was currently uploaded in the supplementary materials as allowed per AAAI 2025 policy.

Evaluation

For a passage retrieval task, we compared the accuracy of the following 8 state-of-the art passage retrieval methods: BM25 (Robertson, Zaragoza et al. 2009), DPR (Karpukhin et al. 2020), RAG (Lewis et al. 2020b), KERS (Zhang et al. 2021), DSI (Tay et al. 2022), Contriever (Izacard et al. 2021), CoT-MAE (Wu et al. 2023), and RankGPT (Sun et al. 2023).

For a response generation task, we compared the quality of responses generated by the following 2 CRS methods (KERS (Zhang et al. 2021) and UniMIND (Deng et al. 2023)), and the following 6 language generation models (GPT-base[1] (Radford et al. 2018), GPT-large (Radford et al. 2018), BART-base[2] (Lewis et al. 2020a), BART-large (Lewis et al. 2020a), RAG (Lewis et al. 2020b), and LLaMa2 (Touvron et al. 2023)).

Note that we could not evaluate other CRS methods (*e.g.*, KBRD (Chen et al. 2019), KGSF (Zhou et al. 2020), CRWaler (Ma, Takanobu, and Huang 2021), RevCore (Lu et al. 2021), and UniCRS (Wang et al. 2022)) utilizing DBpedia KG (Lehmann et al. 2015) because entities of DBpedia KG are not linked in datasets we used. For a similar reason, there have been no papers reporting the performance of CRS methods using DBpedia KG on the public CRS dataset DuRecDial2.0.

Additional Experiments

We conducted additional experiments, aiming at answering the following key research questions (RQs):

- **(RQ6)** Does our *labeling approach* produce reliable pseudo-relevant passages?
- **(RQ7)** How sensitive is the adaptive item selection to the threshold θ_{conf}?
- **(RQ8)** How much does the ESPRESSO-enhanced generation module outperform state-of-the-art methods for the response generation task?
- **(RQ9)** How the passage retrieval from ESPRESSO help generate informative REC responses?

(RQ6) Reliability of our labeling approach

We validate the reliability of our labeling approach in creating pseudo-relevant passages. We compared the quality of training labels created by five different passage retrieval models: BM25 (Robertson, Zaragoza et al. 2009), DPR (Karpukhin et al. 2020), Contriever (Izacard et al. 2021), CoT-MAE (Wu et al. 2023), and RankGPT (Sun et al. 2023). Among them, BM25 is a statistical method that analyzes the lexical similarity between the *enhanced response* and each of the passages. RankGPT is a reranking method that leverages GPT-4o (Achiam et al. 2023) as a reranker for the results of BM25. The remaining models (*i.e.*, DPR, Contriever, and CoT-MAE) are neural retrieval models that capture the representation similarity between them. Note that neural retrieval models are used in their pre-trained state for the labeling task, as they can not be fine-tuned due to *the absence of training labels at this stage*.

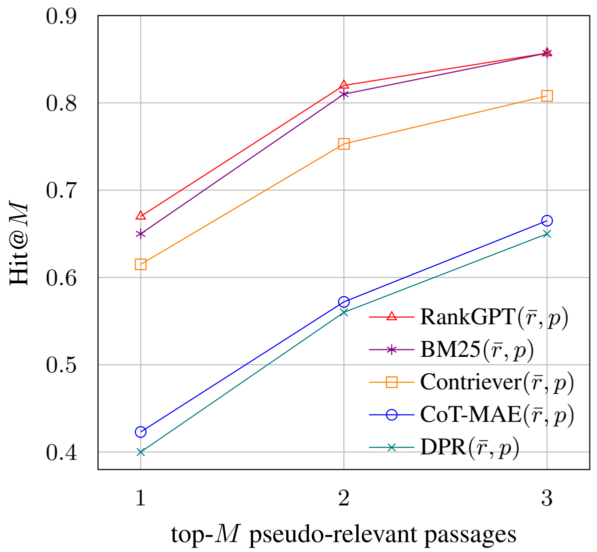

Figure 2: The accuracies of pseudo-relevant passages created by five labeling models in DuRecDial2.0.

Figure 2 shows the hit ratio (*i.e.*, Hit@M) of five different models to evaluate whether their top-M pseudo-relevant passages contain the ground-truth passage where M ranges from 1 to 3. The results indicate that among the five models, RankGPT produces the most reliable pseudo-relevant passages, followed by BM25. These results indicate that our labeling method, which utilizes both a statistical method (*i.e.*, BM25) and a large language model (*i.e.*, GPT-4o), efficiently and effectively identifies pseudo-relevant passages that are truly relevant for generating the recommendation responses. Meanwhile, the result shows that neural retrieval models (*i.e.*, DPR, Contriever, and CoT-MAE) are less effective than BM25 in labeling tasks. We conjecture that their performance limitations primarily stem from out-of-domain scenarios caused by the absence of task-specific fine-tuning.

(RQ7) Sensitivity analysis of θ_{conf} in adaptive item selection

We conduct the sensitivity analysis on the threshold (*i.e.*, σ_{conf}) for *adaptive item selection*. Figure 3 shows the passage retrieval accuracy in Hit@3 when the threshold σ_{conf} is increased from 0% to 100%. First, the number of selected items increases from 0 to 3 as higher σ_{conf} allows more items to be selected. Second, the passage retrieval accuracy *rapidly* increases as the σ_{conf} increases from 0% to 10%. This suggests that, as claimed in Direction-1, even a small number of selected items gives the passage retrieval module a chance to improve accuracy. Third, the passage retrieval accuracy *steadily* increases as the σ_{conf} increases from 10%

[1] We used KoGPT for KoRecDial dataset.

[2] We used KoBART for KoRecDial dataset.

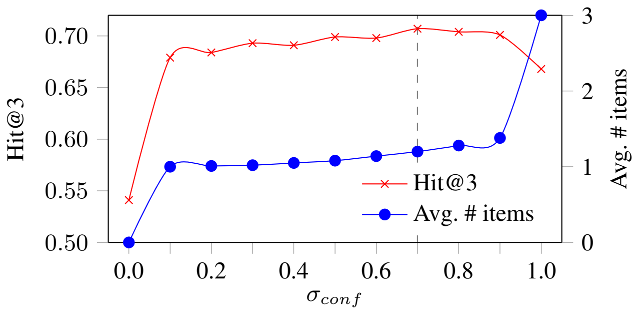

Figure 3: Passage retrieval accuracy at Hit@3 and the number of selected items of ESPRESSO according to different σ_{conf} in adaptive item selection.

Table 1: Comparison of ESPRESSO enhanced generation module and 8 competitors on the response generation task using GPTEval. Here, the best and second-best performers are highlighted by **bold** and underlined, respectively.

Methods	Infor.	Fluency	Relevance
GPT2-base	2.96	3.67	3.20
GPT2-large	3.13	3.79	3.26
BART-base	3.02	3.80	3.22
BART-large	3.13	3.82	3.27
RAG	3.20	3.83	3.32
KERS	2.43	3.27	2.74
UNIMIND	2.81	3.50	2.86
LLaMa2	3.22	3.84	3.40
BART(ESPRESSO)	3.46	3.95	3.66
LLaMa2(ESPRESSO)	**3.53**	**3.98**	**3.72**

to 70%. This indicates that as σ_{conf} increases, the chance of selecting the item that matches the true user preference increases. Finally, the retrieval accuracy decreases when the σ_{conf} increases from 70% to 100%. This indicates that setting too high σ_{conf} could risk selecting noisy items which would harm the passage retrieval performance.

(RQ8) Comparision with competitors for response generation using GPTEval

To further assess the quality of the generated responses, we conducted an additional experiment inspired by GPTEval (Liu et al. 2023). In this experiment, GPT-4o (Achiam et al. 2023), instead of human workers, rated each generation module's responses on a 1-5 scale in terms of *informativeness* (Zhou et al. 2020), *fluency* (Ma, Takanobu, and Huang 2021), and *relevance* (Kim et al. 2023).

Table 1 shows the results for our methods, BART(ESPRESSO) and LLaMa(ESPRESSO), and 8 competitors: GPT2-base, GPT2-large (Radford et al. 2018), BART-base, BART-large (Lewis et al. 2020a), RAG (Lewis et al. 2020b), KERS(Zhang et al. 2021), UniMIND (Deng et al. 2023), and LLaMa2 (Touvron et al. 2023). Notably, BART(ESPRESSO) (resp. LLaMa(ESPRESSO)) achieved improvements of 7.5% (resp. 9.6%) in informativeness, 2.9% (resp. 3.6%) in fluency, and 7.6% (resp. 9.4%) in relevance compared to the best competitor, LLaMa2. This suggests that our accurate passage retrieval strategy significantly enhances the quality of responses generated by the CRS.

CONTEXT

User: Hello, who starred in Bruce Lee: A Warrior's Journey?

System: Jackie Chan.

User : OK, thank you. You're great!

⋮

System: He's outstanding! He has won the 89th Academy Honorary Award for Lifelong Achievement.

User : It's impressive! I admire him even more.

RETRIEVED PASSAGES

1. The Foreigner **stars Jackie Chan**, Pierce Brosnan, and Liu Tao.

2. The Foreigner is known as the best movie of Jackie Chan in recent years.

3. The Foreigner is known as its **awesome plot and captivating story.**

RESPONSE

BART: I recommend **The Foreigner**. It's a good movie

BART(ESPRESSO): I highly recommend watching **The Foreigner**, **starring Jackie Chan**. It's a fantastic movie with a **captivating plot**.

Figure 4: Case study of generated results from BART and BART(ESPRESSO). Here, 'CONTEXT' indicates a sample dialog, 'RETRIEVED PASSAGES' indicates top-3 retrieved passages from ESPRESSO, and 'RESPONSE' indicates the generated results from the above two generative models.

(RQ9) Case study

We present a real-world example to show that retrieved passages from ESPRESSO indeed help generate informative and factually-correct REC-responses. Figure 4 shows a dialog sample with retrieved top-K passages from ESPRESSO along with responses from BART and BART(ESPRESSO). In Figure 4, a user exhibits her preference for 'Jackie Chan'. In this situation, BART generates a response by recommending the item to the user with a trivial explanation. We hypothesize that this issue arises when the internal knowledge for the recommended item in BART is lacking. On the other hand, it is worth noting that BART(ESPRESSO) generates a high-quality REC-response by referring to the well-retrieved passages. Specifically, ESPRESSO successfully retrieves relevant passages that include factually-correct information related to the recommended movie, such as the starring actor (*i.e.*, 'Jackie Chan'), and impressions (*i.e.*, "awesome plot and captivating story"). This confirms that, in CRS, to generate REC-responses in a factually-correct manner, a well-designed passage retrieval method like ESPRESSO is essential.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, Q.; Lin, J.; Zhang, Y.; Ding, M.; Cen, Y.; Yang, H.; and Tang, J. 2019. Towards Knowledge-Based Recommender Dialog System. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1803–1813.

Deng, Y.; Zhang, W.; Xu, W.; Lei, W.; Chua, T.-S.; and Lam, W. 2023. A Unified Multi-task Learning Framework for Multi-goal Conversational Recommender Systems. *ACM Transactions on Information Systems*, 41(3): 1–25.

Hayati, S. A.; Kang, D.; Zhu, Q.; Shi, W.; and Yu, Z. 2020. Inspired: Toward Sociable Recommendation Dialog Systems. *arXiv preprint arXiv:2009.14306*.

Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2021. Unsupervised Dense Information Retrieval with Contrastive Learning. *arXiv preprint arXiv:2112.09118*.

Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2004.04906*.

Kim, T.; Yu, J.; Shin, W.-Y.; Lee, H.; Im, J.-h.; and Kim, S.-W. 2023. LATTE: A Framework for Learning Item-Features to Make a Domain-Expert for Effective Conversational Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1144–1153.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; et al. 2015. DBpedia–A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic web*, 6(2): 167–195.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020a. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020b. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, R.; Ebrahimi Kahou, S.; Schulz, H.; Michalski, V.; Charlin, L.; and Pal, C. 2018. Towards Deep Conversational Recommendations. *Advances in Neural Information Processing Systems*, 31.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. Gpteval: Nlg Evaluation Using GPT-4 with Better Human Alignment. *arXiv preprint arXiv:2303.16634*.

Liu, Z.; Wang, H.; Niu, Z.-Y.; Wu, H.; and Che, W. 2021. DuRecDial 2.0: A Bilingual Parallel Corpus for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4335–4347.

Lu, Y.; Bao, J.; Song, Y.; Ma, Z.; Cui, S.; Wu, Y.; and He, X. 2021. RevCore: Review-augmented conversational recommendation. *arXiv preprint arXiv:2106.00957*.

Ma, W.; Takanobu, R.; and Huang, M. 2021. CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1839–1851.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving Language Understanding by Generative Pre-Training.

Robertson, S.; Zaragoza, H.; et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Sun, W.; Yan, L.; Ma, X.; Wang, S.; Ren, P.; Chen, Z.; Yin, D.; and Ren, Z. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*.

Tay, Y.; Tran, V.; Dehghani, M.; Ni, J.; Bahri, D.; Mehta, H.; Qin, Z.; Hui, K.; Zhao, Z.; Gupta, J.; et al. 2022. Transformer Memory as a Differentiable Search Index. *Advances in Neural Information Processing Systems*, 35: 21831–21843.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. LLaMa 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Wang, X.; Zhou, K.; Wen, J.-R.; and Zhao, W. X. 2022. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1929–1937.

Wu, X.; Ma, G.; Lin, M.; Lin, Z.; Wang, Z.; and Hu, S. 2023. ConTextual Masked Auto-Encoder for Dense Passage Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4738–4746.

Zhang, J.; Yang, Y.; Chen, C.; He, L.; and Yu, Z. 2021. KERS: A knowledge-Enhanced Framework for Recommendation Dialog Systems with Multiple Subgoals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1092–1101.

Zhou, K.; Zhao, W. X.; Bian, S.; Zhou, Y.; Wen, J.-R.; and Yu, J. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1006–1014.