Aspect-based Evaluation of Personalization and Diversification in Conversational Search Systems

Risa Hiramatsu[1,*], Rikiya Takehi[1] and Tetsuya Sakai[1]

[1]Waseda University, Tokyo, Japan

Abstract

Recent advances in conversational search systems and AI-powered agents have highlighted the importance of personalization, as providing customized responses for each user is key to meeting individual needs. At the same time, diversification, which can be considered the opposite of personalization, is equally important to provide various information regardless of the user. To meet the various needs of users who may prefer a personalized system or a diversified system, it is desirable to enable the evaluation of both simultaneously. For assessing personalization, most previous work has measured it by checking whether the contents mentioned in the user's personal profile are covered in the conversation. This is problematic because it can overrate conversations that simply use many profile words and ignore the genuine relevance between the conversation content and the profile. Moreover, to the best of our knowledge, there is no existing framework that unifies the evaluation of personalization and diversification by placing them at opposite ends of a spectrum.

We propose **A**spect-based **E**valuation of **P**ersonalization and **D**iversification (**AEPD**), an evaluation framework for conversational search systems such as those based on RAG (retrieval-augmented generation) that assesses conversational turns at the aspect level. We define aspects as the various facets or answers for the topic given in the conversation. In terms of personalization, this framework assesses whether the aspects raised in conversational turns are relevant to the personal profile, thus making it possible to place importance on the actual contents of the system response. It also evaluates diversification in terms of unique aspects raised. This framework is also novel in that it quantifies both personalization and diversification simultaneously in conversational search systems. We conduct a pilot experiment on the TREC iKAT 2024 dataset leveraging ChatGPT o4-mini, demonstrating how AEPD can be computed automatically as well as the personalization-diversification trade-off.

Keywords

Retrieval–Augmented Generation, Large Language Models, Evaluation, Information Retrieval, Aspects, Personalization, Diversification

1. Introduction

Generative conversational systems have made substantial progress in recent years [1, 2, 3], and so many evaluation metrics have been proposed to keep up [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Some evaluate system responses based on either how personalized [17, 18, 19, 20] or diversified [21, 22, 23, 24] they are, aiming to better meet users' information needs.

In terms of personalization, most existing metrics focus on the consistency with the user's given personal profile [25, 26, 27]. However, our view is that existing measures are not adequate for evaluating whether system responses are personalized or not. For example, in Figure 1, the system response is likely to be evaluated as personalized just because it greets the user as "information retrieval expert" in reference to the "IR researcher" statement in the user's personal profile. However, we do not want to take "information retrieval expert" into consideration because it is not relevant to the given topic "Things to do in free time". Instead, we want to consider this system response poor in terms of personalization because the answer for the topic "How about listening to music?" does not align with the statement "I like to read books" in the user profile.

Moreover, no previous metrics have seamlessly evaluated personalization and diversification along the same spectrum. If this is possible, it may provide a starting point for establishing a foundation for

This paper was reviewed and accepted by the program committee for BREV-RAG 2025: Beyond Relevance-based EValuation of RAG Systems, a SIGIR-AP 2025 workshop, held on December 10, 2025 in Xi'an, China.

*Corresponding author.

risahira@fuji.waseda.jp (R. Hiramatsu); rikiya.takehi@fuji.waseda.jp (R. Takehi); tetsuyasakai@acm.org (T. Sakai)

0009-0005-9995-8458 (R. Hiramatsu); 0009-0003-6336-8064 (R. Takehi); 0000-0002-6720-963X (T. Sakai)

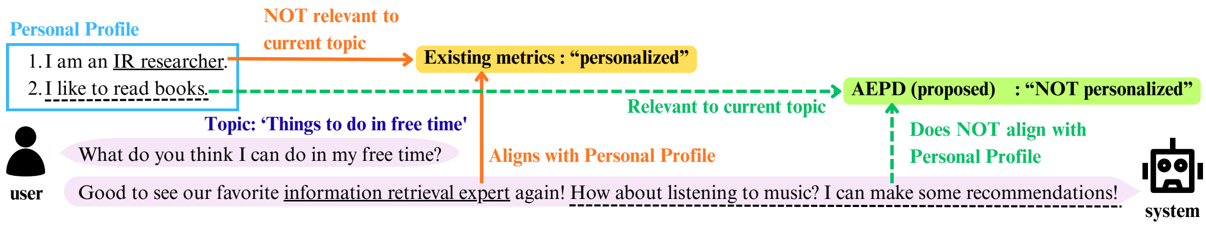


Figure 1: Comparison between our AEPD and previous work(C score [25], P-Score [26], Con.Score [27]) on how personalization would be evaluated.

users to adjust the balance between the two according to their preferences, given that the two can be regarded as being in an opposing relationship. We argue that enabling this adjustment is important because some users prefer personalized responses that adhere to their own context, while others prefer diversified responses that broaden their perspectives.

To solve the issues above, we propose **A**spect-based **E**valuation of **P**ersonalization and **D**iversification (**AEPD**), an evaluation framework that simultaneously measures the degree of personalization and diversification in conversational search systems. In particular, we focus on the *aspects* that are given for each topic in a conversational turn. We define *aspects* as the various facets that a given topic contains (also known as *subtopics* or *intents* [28]) as well as possible answers relevant to the topic. For example, in Figure 1, the topic is "Things to do in free time", and possible existing aspects for the topic would be: listening to music, reading books, going for a walk, etc. By focusing on *aspects*, we can appropriately evaluate the genuine content of the system response. Unlike most of the prior metrics, we evaluate personalization based on the aspects that appear in the conversational turn *and* are relevant to the personal profile. Thus, we can avoid the problem mentioned above and evaluate the system in Figure 1 as non-personalized.

Our framework proposes novel concepts of a *personalized* and *diversified states* in conversational search systems. A *personalized state* is defined as "the response provides sufficient information for specific aspects that suits the user about the given topic", whereas a *diversified state* is defined as "the response provides a uniform depth of information for diverse aspects about the given topic". When the system is highly personalized, it discusses an aspect that is relevant to the user's personal profile in detail, making it difficult to cover other aspects. Conversely, when the system is highly diversified, it covers various aspects in a balanced manner, making it difficult to go deeper into any single aspect. Therefore, a trade-off exists between a personalized state and a diversified state.

Our contributions are threefold: (1) highlighting the importance of adjusting the balance between personalization and diversification, (2) enabling seamless evaluation of these two features, and (3) evaluating both features by the genuine contents of the system response. Moreover, by conducting a pilot experiment in which ChatGPT o4-mini automatically computes AEPD, we show how personalization and diversification can be quantified.

2. Related Work

2.1. Evaluation of Personalization in Conversations

Although we have so far not found any metric that evaluates both personalization and diversification in conversational search systems, a query suggestion paradigm called QS-DP [29] integrates diversification and personalization in query suggestion. However, this method first diversifies the query suggestion candidates, then personalizes the rankings of the candidates. Thus, personalization and diversification are sequentially combined and there is no trade-off here, and therefore users cannot adjust the balance to their preference.

Chen et al. [30] referred to several evaluation metrics on the personalization of dialogue generation, and many have evaluated dialogues based on the consistency with the given personal profile [25, 26, 27].

Table 1
Notations for AEPD

Notation	Description
T	A specific conversational turn.
t	The topic introduced during Turn T.
u	The user who participated in Turn T.
$A_{\text{all}}(T,t)$	Set of aspects that appeared in Turn T for Topic t.
$A_{\text{per}}(T,t,u)$	$\subseteq A_{\text{all}}(T,t)$ Set of personalized aspects for User u.
$A_{\text{nonper}}(T,t,u)$	$= A_{\text{all}}(T,t) - A_{\text{per}}(T,t,u)$ Set of non-personalized aspects for User u.
$\mathcal{A}'_{\text{all}}(t)$	Set of possible aspects that could exist for Topic t in general. (Topic t has other possible aspects not mentioned in Turn T.)
$\mathcal{A}'_{\text{per}}(t,u)$	$\subseteq \mathcal{A}'_{\text{all}}(t)$ Set of possible personalized aspects for User u.
$l_{\text{all}}(T,t,a)$	Length (number of words) of content that is relevant to Aspect $a \in A_{\text{all}}(T,t)$.
$l_{\text{per}}(T,t,a,u)$	Length (number of words) of content that is relevant to Aspect $a \in A_{\text{per}}(T,t,u)$.
$l_{\text{nonper}}(T,t,a,u)$	Length (number of words) of content that is relevant to Aspect $a \in A_{\text{nonper}}(T,t,u)$.

Some metrics using LLMs have been proposed [31, 32], but do not perform aspect-based evaluations. Gerritse et al. [33] argued that conversational search could create biases through personalization. Their idea is similar to ours in that it recognizes a trade-off relationship between a personalized state and a non-biased state (i.e., diversified state), but they focus only on detecting biases.

2.2. Evaluation of Diversification in Conversations

Li et al. [34] evaluated the degree of diversity by the number of distinct unigrams and bigrams, but this approach cannot consider diversity in terms of aspects under the current topic. Han et al. [35] measured the semantic diversity, but did not evaluate whether each aspect is presented with balanced lengths. The same can be said of Guo et al. [36], who proposed a similar concept as our work, but does not measure the balance across diverse aspects in terms of textual length. Other metrics for diversity also exist [37, 38, 39], but none of them measure diversification based on aspects.

3. Aspect-based Evaluation of Personalization and Diversification (AEPD)

3.1. Problem Setting

In a conversational search session, a user and a system discuss one or more topics. Throughout this paper, we refer to a conversational search session as a *conversation*, and one exchange by the user and the system (which is a part of a conversation) as a *conversational turn*. We show the notations we use (in Section 3.2 and Section 3.3) in Table 1, and an example of a conversational turn between a user and a system in Figure 2. In Table 1, "Personalized" means the content is considered relevant to the user's personal profile, and "length" means the number of words. In this framework, we assume that the user's personal profile is provided to the system beforehand. A user's personal profile may include user context or past interactions.

Our approach calculates *Personalization Score* and *Diversification Score* for each conversational turn. Using the notations described in Table 1, Figure 2 also shows the values that are needed for the calculations in Section 3.2 and Section 3.3.

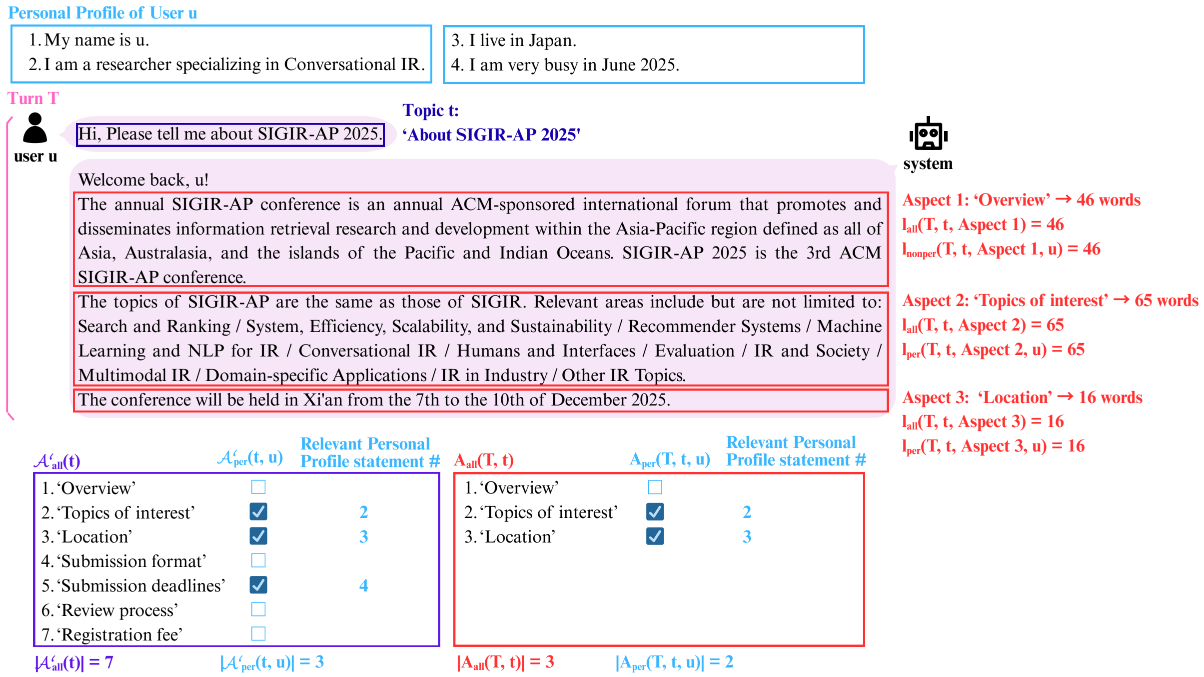


Figure 2: An example of a conversational turn between a user and a system. The user's Personal Profile is given in the top blue box, and Turn T is given below. The left bottom purple box shows there are 7 existing aspects for topic t. The right bottom red box shows that 3 aspects appeared in the conversational turn. Aspects that are relevant to the Personal Profile have check marks beside them.

3.2. Personalization Score

A personalized system response should contain aspects that are relevant to the user's personal profile, and discuss more about those aspects compared to other non-relevant aspects. We aim to measure the former in Section 3.2.1, and the latter in Section 3.2.2.

3.2.1. Normalized Personalized Precision

We first define the **P**ersonalized **P**recision for the conversational turn and in general. The Personalized Precision inside the conversational turn is given by:

$$PP(T, t, u) = \frac{|A_{\text{per}}(T, t, u)|}{|A_{\text{all}}(T, t)|} \tag{1}$$

$|A_{\text{all}}(T, t)| \geq |A_{\text{per}}(T, t, u)|$ always stands, so $PP(T, t, u) \in [0, 1]$. In Figure 2, the red boxes show the system response includes 3 aspects, thus $|A_{\text{all}}(T, t)| = 3$. In addition, Aspect 2 "Topics of interest" is relevant to Personal Profile statement 2 "I am a researcher specializing in Conversational IR." because it shows User u's expertise in a specific topic, and Aspect 3 "Location" is relevant to Personal Profile statement 3 "I live in Japan." because both are about locations. Thus, $|A_{\text{per}}(T, t, u)| = 2$.

The Personalized Precision in general is given by:

$$\mathcal{PP}(t, u) = \frac{|\mathcal{A}'_{\text{per}}(t, u)|}{|\mathcal{A}'_{\text{all}}(t)|} \tag{2}$$

$|\mathcal{A}'_{\text{all}}(t)| \geq |\mathcal{A}'_{\text{per}}(t, u)|$ always stands, so $\mathcal{PP}(t, u) \in [0, 1]$. In Figure 2, the bottom left purple box shows there are 7 aspects for Topic t in general, thus $|\mathcal{A}'_{\text{all}}(t)| = 7$. Moreover, following the previous explanation of $A_{\text{per}}(T, t, u)$, Aspects 2 and 3 are relevant to the personal profile. In addition, Aspect 5 "Submission deadlines" is relevant to Personal Profile statement 4 "I am very busy in June 2025" because deadlines are important information for a busy person. Thus, $|\mathcal{A}'_{\text{per}}(t, u)| = 3$.

Then, we compare the proportion of personalized aspects in the conversational turn to the proportion of personalized aspects among all possible aspects of topic t. This is based on our hypothesis that a personalized system should provide a higher concentration of personalized aspects compared to existing aspects in general. Therefore, $\mathcal{PP}(t, u)$ will be the baseline for comparison. Here, we introduce **N**ormalized **P**ersonalized **P**recision (NPP), which measures how well personalized aspects were selected in the response. It rewards conversational turns that are more personalized than the baseline. The formula adopts a similar normalization structure as the Kappa Coefficient, and incorporates ReLU to prevent negative values:

$$NPP(T, t, u) = \frac{\max\{0, PP(T, t, u) - \mathcal{PP}(t, u)\}}{1 - \mathcal{PP}(t, u)} \tag{3}$$

where we define $NPP(T, t, u) = 0$ if $\mathcal{PP}(t, u) = 1$. $PP(T, t, u) \leq 1$ is always satisfied, so $NPP(T, t, u) \in [0, 1]$.

3.2.2. Normalized Personalized Content Length

We define $L_{\text{all}}(T, t)$ as the total length of contents of all given aspects in Turn T for Topic t in the system response, and is given by:

$$L_{\text{all}}(T, t) = \sum_{a \in A_{\text{all}}(T,t)} l_{\text{all}}(T, t, a) \tag{4}$$

In Figure 2, the red sentences beside the system response shows Aspect 1 contains 46 words, so $l_{\text{all}}(T, t, Aspect\ 1) = 46$. In the same way, $l_{\text{all}}(T, t, Aspect\ 2) = 65$, $l_{\text{all}}(T, t, Aspect\ 3) = 16$.

For each $a \in A_{\text{per}}(T, t, u)$, we define the average of $l_{\text{per}}(T, t, a, u)$ as:

$$L_{\text{per}}(T, t, u) = \frac{\sum_{a \in A_{\text{per}}(T,t,u)} l_{\text{per}}(T, t, a, u)}{|A_{\text{per}}(T, t, u)|} \tag{5}$$

where we define $L_{\text{per}}(T, t, u) = 0$ if $A_{\text{per}}(T, t, u) = \emptyset$. In Figure 2, Aspect 2 and Aspect 3 are personalized in the response, so $l_{\text{per}}(T, t, Aspect\ 2) = 65$, $l_{\text{per}}(T, t, Aspect\ 3) = 16$, and $|A_{\text{per}}(T, t, u)| = 2$.

For each $a \in A_{\text{nonper}}(T, t, u)$, we define the average of $l_{\text{nonper}}(T, t, a, u)$ as:

$$L_{\text{nonper}}(T, t, u) = \frac{\sum_{a \in A_{\text{nonper}}(T,t,u)} l_{\text{nonper}}(T, t, a, u)}{|A_{\text{nonper}}(T, t, u)|} \tag{6}$$

where we define $L_{\text{nonper}}(T, t, u) = 0$ if $A_{\text{nonper}}(T, t, u) = \emptyset$. In Figure 2, Aspect 1 is not personalized, so $l_{\text{nonper}}(T, t, Aspect\ 1) = 46$ and $|A_{\text{nonper}}(T, t, u)| = 1$.

Next, we compare the content length of personalized aspects to the content length of non-personalized aspects. This is based on our hypothesis that a personalized system should allocate more content to personalized aspects compared to non-personalized aspects. To measure this, we introduce **N**ormalized **P**ersonalized **C**ontent **L**ength ($NPCL$), which quantifies how deeply personalized aspects were discussed in the conversational turn. Here, the depth is quantified by the number of words of the content. We again use the same formula structure as Eq. (3) and reward a conversational turn if personalized contents contain more words compared to non-personalized contents:

$$NPCL(T, t, u) = \frac{\max\{0, L_{\text{per}}(T, t, u) - L_{\text{nonper}}(T, t, u)\}}{L_{\text{all}}(T, t) - L_{\text{nonper}}(T, t, u)} \tag{7}$$

where we define $NPCL(T, t, u) = 0$ if $L_{\text{all}}(T, t) = L_{\text{nonper}}(T, t, u)$. Note that $NPCL(T, t, u) \in [0, 1]$, since $L_{\text{all}}(T, t) \geq L_{\text{per}}(T, t, u)$.

3.2.3. Personalization Score

Finally, to gain the Personalization Score, we consider using NPP and $NPCL$ as the two factors that measure how personalized a system response is. We aim to evaluate both factors jointly, such that if either one is close to zero, the overall score will also be low. Thus, we multiply the two factors, and define the **P**ersonalization Score as:

$$P_{\text{score}}(T, t, u) = NPP(T, t, u) \times NPCL(T, t, u) \tag{8}$$

$P_{\text{score}}(T, t, u) \in [0, 1]$, and the more personalized a system is in the conversational turn, the higher the score.

3.3. Diversification Score

A diversified system response should provide as many aspects regardless of any user, and the length of content for each aspect should be balanced to ensure fair representation of diverse information. We aim to measure the former in Section 3.3.1, and the latter in Section 3.3.2.

3.3.1. Aspect Recall

To measure the coverage of aspects in the system response, we introduce **A**spect **R**ecall (AR), which is defined by:

$$AR(T, t) = \frac{|A_{\text{all}}(T, t)|}{|\mathcal{A}'_{\text{all}}(t)|} \tag{9}$$

Since $|\mathcal{A}'_{\text{all}}(t)| \geq |A_{\text{all}}(T, t)|$ always stands, $AR \in [0, 1]$. In Figure 2, $|A_{\text{all}}(T, t)| = 3$, $|\mathcal{A}'_{\text{all}}(t)| = 7$.

3.3.2. Content Length Uniformity

In this section, our goal is to calculate the uniformity of the number of words of each aspect content that appeared in the conversational turn.

A normalized content length of aspect $a \in A_{\text{all}}(T, t)$ will be:

$$l'_{\text{all}}(T, t, a) = \frac{l_{\text{all}}(T, t, a)}{L_{\text{all}}(T, t)} \tag{10}$$

where $L_{\text{all}}(T, t)$ is from Eq. (4). In Figure 2, $l_{\text{all}}(T, t, Aspect\ 1) = 46$, $l_{\text{all}}(T, t, Aspect\ 2) = 65$, $l_{\text{all}}(T, t, Aspect\ 3) = 16$.

Moreover, we define $l^*(t)$, which represents the uniform distribution of $l'_{\text{all}}(T, t, a)$, as:

$$l^*_{\text{all}}(T, t) = \frac{1}{|A_{\text{all}}(T, t)|} \tag{11}$$

In Figure 2, $|A_{\text{all}}(T, t)| = 3$, so $l^*_{\text{all}}(T, t) = \frac{1}{3}$.

Then, we calculate the *Root Normalized Sum of Squares* ($RNSS$) [40] which represents the deviation from a uniform distribution. $RNSS$ is large when uniformity is low.

$$SS(l'_{\text{all}}(T, t, a), l^*_{\text{all}}(T, t)) = \sum_{a \in A_{\text{all}}(T, t)} \left(l'_{\text{all}}(T, t, a) - l^*_{\text{all}}(T, t)\right)^2 \tag{12}$$

$$RNSS(l'_{\text{all}}(T, t, a), l^*_{\text{all}}(T, t)) = \sqrt{\frac{SS(l'_{\text{all}}(T, t, a), l^*_{\text{all}}(T, t))}{2}} \tag{13}$$

Finally, we calculate **C**ontent **L**ength **U**niformity (CLU), which measures the uniformity of the length of content for each aspect in the conversational turn. We subtract $RNSS$ from 1, which is the max value of $RNSS$. Thus, CLU is low when the lengths of the content are not uniform.

$$CLU(T, t) = 1 - RNSS(l'_{\text{all}}(T, t, a), l^*_{\text{all}}(T, t)) \tag{14}$$

Since $RNSS(l'_{\text{all}}(T, t, a), l^*_{\text{all}}(T, t)) \in [0, 1]$, $CLU(T, t) \in [0, 1]$ also stands.

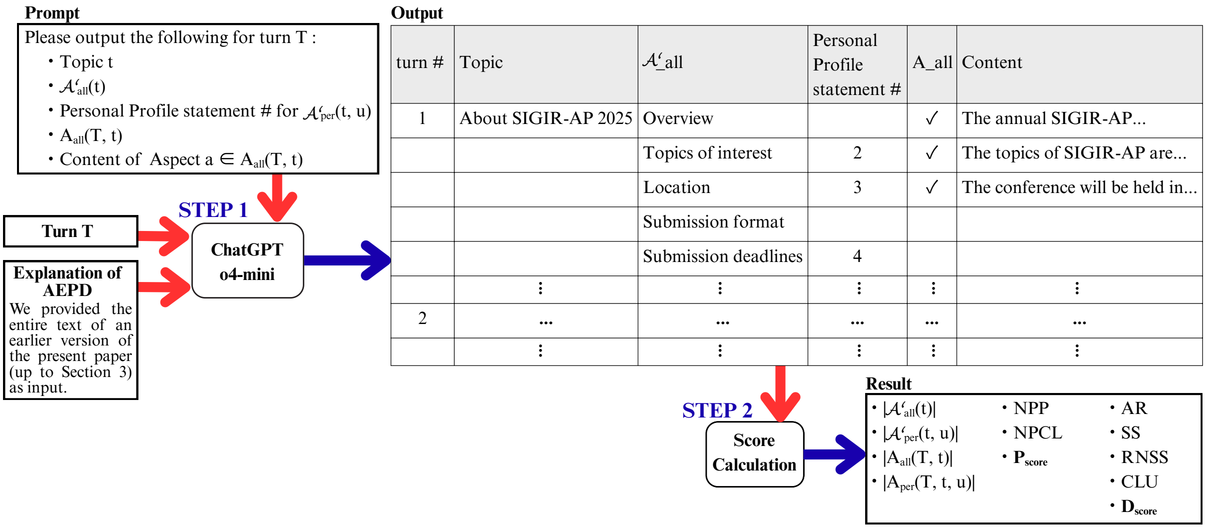


Figure 3: An example flow of our AEPD framework. In STEP 1, an instruction prompt is given to ChatGPT o4-mini to generate information needed to calculate the scores. We gained an output for each conversational turn. The check marks in the A_{all} column of the output show which aspects appeared in the conversational turn. In STEP 2, the scores are calculated.

3.3.3. Diversification Score

To compute the final Diversification Score, we use AR and CLU as the two factors that measure how diversified a system response is in the conversational turn. The more it provides diversified contents, the higher the score should be. As in Section 3.2.3, here we again multiply AR and CLU for the same reason. We define the **D**iversification Score as:

$$D_{\text{score}}(T, t) = AR(T, t) \times CLU(T, t) \tag{15}$$

where $D_{\text{score}}(T, t) \in [0, 1]$.

4. Pilot Experiment

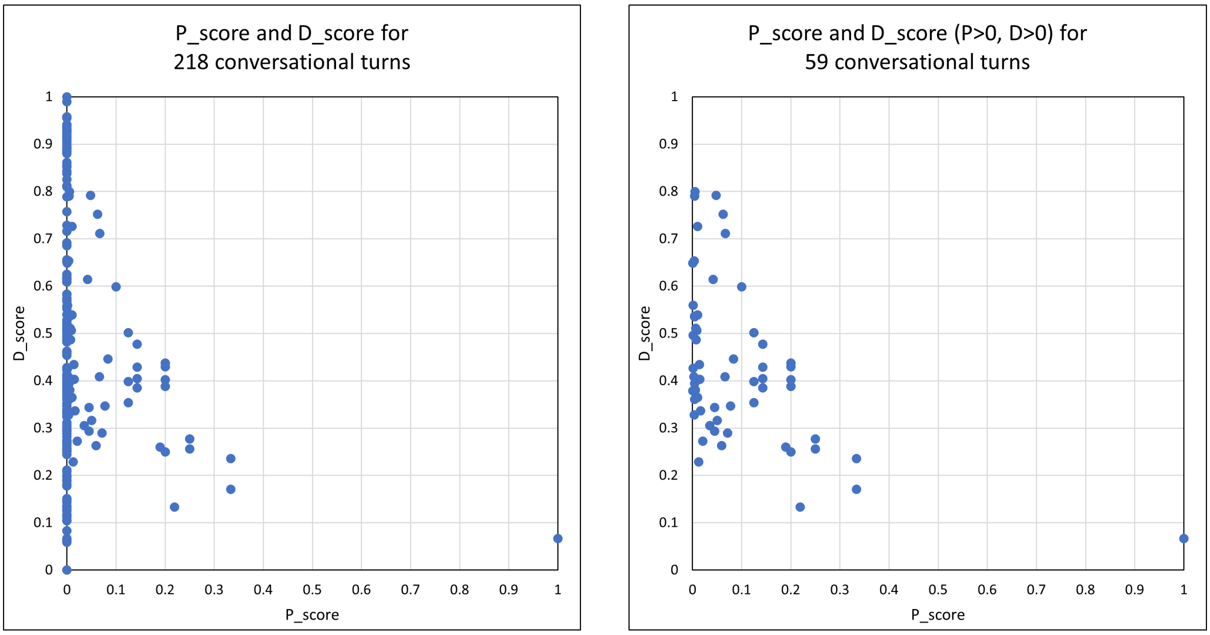
4.1. Setup

To demonstrate the feasibility of our AEPD framework, we conducted a pilot experiment that leverages ChatGPT o4-mini[1] for computing our measures.

The dataset of conversations we used was the TREC iKAT 2024 collection [41], which contains 17 conversations with multiple turns and Personal Text Knowledge Base (PTKB) statements for each. We considered "`number`" as the conversation number, "`ptkb`" as the user's personal profile, "`turns`" as conversational turns, "`turn_id`" as the conversational turn number, "`utterance`" as user utterance, and "`response`" as system response.

We show an example flow of the experiment in Figure 3. In STEP 1, we provide ChatGPT with a conversational turn, an explanation of AEPD, and a prompt[2]. ChatGPT is instructed to generate the topic of each turn and the existing aspects for it. The topics and aspects are generated automatically by the LLM, and do not require manual annotation. Then, it checks whether the aspects align with the personal profile and whether the aspects appeared in the turn. Finally, it extracts the parts of the content relevant to the aspects. In STEP 2, we compute the final scores[2] based on the output from STEP 1.

[1] https://chatgpt.com/
[2] Our prompts and code are available at https://github.com/risa-hira/AEPD



(a) Results of all 218 conversational turns in TREC iKAT 2024.

(b) Results of 59 conversational turns with $P_{\text{score}} > 0$ and $D_{\text{score}} > 0$, extracted from Figure 4a. Kendall's τ: $\tau = \mathbf{-0.328} < \mathbf{0}$, p-value: $\mathbf{p} = \mathbf{0.0003}$

Figure 4: Results of the assessment of 17 conversations, 218 conversational turns in TREC iKAT 2024. The plot presents P_{score} on the horizontal axis and D_{score} on the vertical axis.

4.2. Results

The results of the LLM assessment of a total of 17 conversations, 218 conversational turns, are shown in Figure 4. Figure 4a shows the P_{score} and D_{score} for all 218 conversational turns of TREC iKAT 2024. This suggests that conversational systems can be quantitatively evaluated to determine whether they provide personalized or diversified content.

As we have defined in Section 1, a system cannot be highly personalized and highly diversified simultaneously, since it is impossible for the system to be personalized to the user and give sufficient information for such personalized aspects, and at the same time provide a unified amount of information for diverse aspects. This generates a trade-off between P_{score} and D_{score}. To highlight this trade-off relationship, we have extracted results with $P_{\text{score}} > 0$ and $D_{\text{score}} > 0$ from Figure 4a, which is shown in Figure 4b, since zero scores do not meaningfully reflect the trade-off.

For Figure 4b, the Kendall Rank Correlation Coefficient was $\tau = \mathbf{-0.328} < \mathbf{0}$, and the p-value was $\mathbf{p} = \mathbf{0.0003} < \mathbf{0.05}$ (n = 59), indicating a statistically significant monotonically decreasing trend, which reveals a trade-off relationship between P_{score} and D_{score}. It highlights the potential to evaluate conversational search systems by treating personalization and diversification as opposite ends of a single spectrum, and assessing each system based on its position along that axis.

4.3. Case Study Based on a Selected Result

In this section, we show an example of how the actual calculations were performed in the experiment, using the results from one of the conversational turns in the TREC iKAT 2024 collection. We have selected `"number": 2, "turn_id": 14`, as it was suitable for explanation in that both values of P_{score} and D_{score} were above 0.

We show the actual assessment process in Figure 5, and the values for calculation in Figure 6. By the assessment in Figure 5 and Figure 6, the calculated result was $P_{\text{score}} = 0.0505$ and $D_{\text{score}} = 0.3165$.

Let us now examine the difference between our AEPD framework and the official results from the

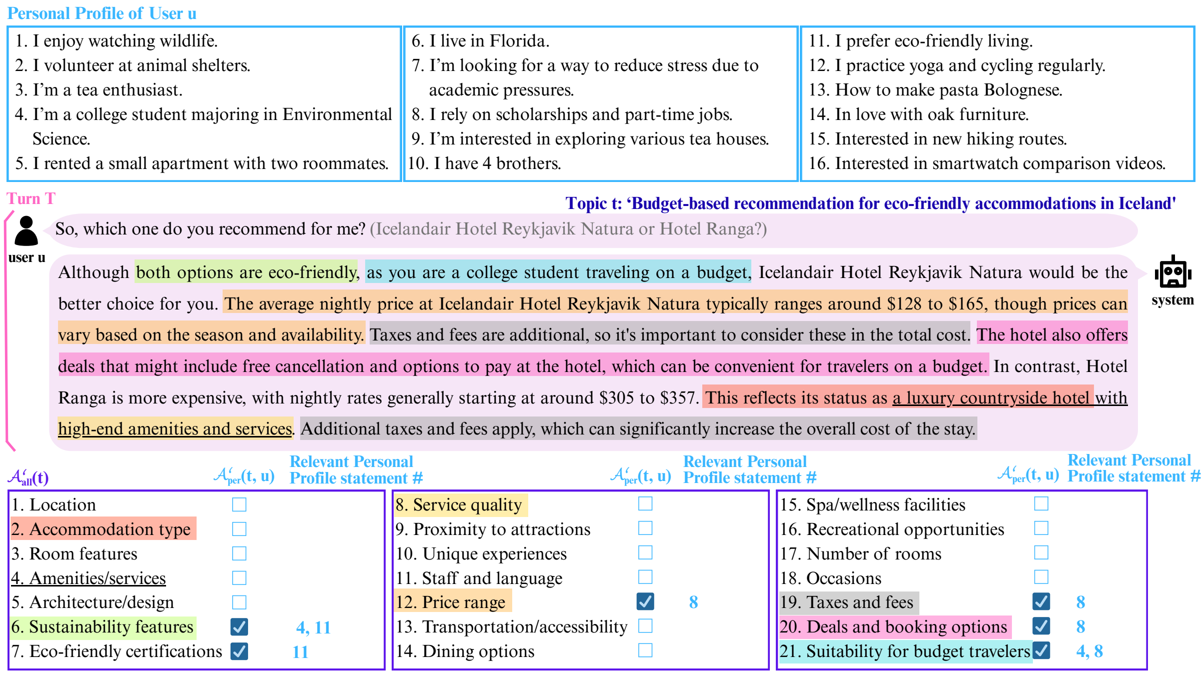


Figure 5: A case study on the TREC iKAT 2024 collection (`"number":2, "turn_id":14`), showing the situation and process of the LLM assessment.

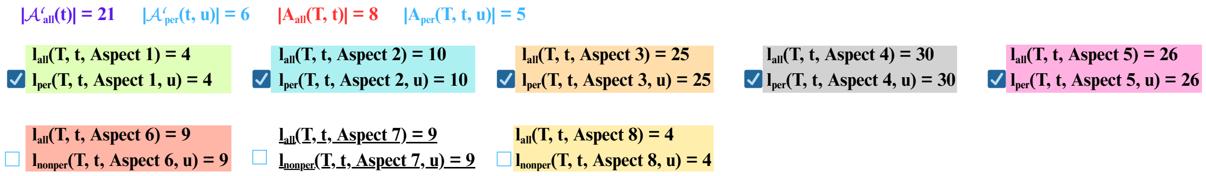


Figure 6: A case study on the TREC iKAT 2024 collection (`"number":2, "turn_id":14`), showing the values used for calculating P_{score} and D_{score}.

TREC iKAT 2024 collection in terms of PTKB assessment. The original PTKB statement relevance assessment was made by the organizers and the NIST assessors in which PTKB statements were classified as relevant or irrelevant for each conversational turn [41]. They concluded that PTKB statements 4 and 11 were relevant in turn 14 of conversation 2. However, the AEPD framework considered that in addition to PTKB statements 4 and 11, PTKB statement 8 was also relevant. This is because PTKB statement 8 (Personal Profile statement 8 in Figure 5) "I rely on scholarships and part-time jobs." may be considered relevant to Aspect 12 "Price range", Aspect 19 "Taxes and fees", Aspect 20 "Deals and booking options", and Aspect 21 "Suitability for budget travelers" because it seems important for the user to save money. This difference between simple relevance judgment and our approach implies the potential of AEPD to recognize contents that are actually relevant to the user, by conducting assessments based on aspects.

5. Conclusion and Future Work

In this paper, we have studied how to evaluate the personalization and diversification of conversational search systems along the same spectrum, and by the genuine contents of the system response. To this end, we proposed AEPD that makes evaluations based on aspects. We believe that this will enable us to evaluate whether the conversational search system is providing the information that the user truly wants. Furthermore, AEPD is also novel in terms of assessing both the ability of personalization and diversification. This may support system developers create benchmarks on the adjustment of the content generated by the system. A conversational search system's capability of adjusting its content is a key to user satisfaction because every user has different information needs, and it depends on the user whether personalization or diversification is more important to them, or they want both features balanced. Furthermore, by conducting a pilot experiment using AEPD, we demonstrated the workability of the framework and the ability to quantify how personalized or diversified a conversational search system is.

Our work has two main limitations. First, the validity of the AEPD assessment has not yet been verified against human judgment. Second, it remains unclear whether the assessment results align with human perceptions of personalization and diversification. Thus, for future work, we plan to compare the AEPD assessment with human assessments, and moreover examine whether the results also align with human perceptions, for example through a user satisfaction survey. In addition, we aim to establish a conversational search system in which the users themselves can adjust the degree of personalization and diversification to their own preference.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to help with grammar and spelling checks, as well as rephrasing sentences or paragraphs to improve clarity and conciseness. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Soudani, R. Petcu, E. Kanoulas, F. Hasibi, A survey on recent advances in conversational data generation, arXiv preprint arXiv:2405.13003 (2024).

[2] J. Ni, T. Young, V. Pandelea, F. Xue, E. Cambria, Recent advances in deep learning based dialogue systems: A systematic survey, Artificial intelligence review 56 (2023) 3055–3155.

[3] A. Algherairy, M. Ahmed, A review of dialogue systems: current trends and future directions, Neural Computing and Applications 36 (2023) 6325–6351. URL: http://dx.doi.org/10.1007/s00521-023-09322-1. doi:10.1007/s00521-023-09322-1.

[4] S. Mehri, M. Eskenazi, Usr: An unsupervised and reference free evaluation metric for dialog generation, arXiv preprint arXiv:2005.00456 (2020).

[5] M. Li, J. Weston, S. Roller, Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons, arXiv preprint arXiv:1909.03087 (2019).

[6] N. Dziri, E. Kamalloo, K. W. Mathewson, O. Zaiane, Evaluating coherence in dialogue systems using entailment, arXiv preprint arXiv:1904.03371 (2019).

[7] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, J. Pineau, Towards an automatic turing test: Learning to evaluate dialogue responses, arXiv preprint arXiv:1708.07149 (2017).

[8] C. Tao, L. Mou, D. Zhao, R. Yan, Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[9] V. Phy, Y. Zhao, A. Aizawa, Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems, arXiv preprint arXiv:2011.00483 (2020).

[10] T. Lan, X.-L. Mao, W. Wei, X. Gao, H. Huang, Pone: A novel automatic evaluation metric for open-domain generative dialogue systems, ACM Transactions on Information Systems (TOIS) 39 (2020) 1–37.

[11] L. Huang, Z. Ye, J. Qin, L. Lin, X. Liang, Grade: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems, arXiv preprint arXiv:2010.03994 (2020).

[12] K. Sinha, P. Parthasarathi, J. Wang, R. Lowe, W. L. Hamilton, J. Pineau, Learning an unreferenced metric for online dialogue evaluation, arXiv preprint arXiv:2005.00583 (2020).

[13] A. B. Sai, A. K. Mohankumar, S. Arora, M. M. Khapra, Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining, Transactions of the Association for Computational Linguistics 8 (2020) 810–827.
[14] S. Mehri, M. Eskenazi, Unsupervised evaluation of interactive dialog with dialogpt, arXiv preprint arXiv:2006.12719 (2020).
[15] S. Ghazarian, N. Wen, A. Galstyan, N. Peng, Deam: Dialogue coherence evaluation using amr-based semantic manipulations, arXiv preprint arXiv:2203.09711 (2022).
[16] S. Ghazarian, R. Weischedel, A. Galstyan, N. Peng, Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7789–7796.
[17] B. Li, H. Deng, Bilateral personalized dialogue generation with contrastive learning, Soft Computing 27 (2023) 3115–3132.
[18] C. Miyazaki, S. Kanno, M. Yoda, J. Ono, H. Wakaki, Fundamental exploration of evaluation metrics for persona characteristics of text utterances, in: Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue, 2021, pp. 178–189.
[19] V. Samuel, H. P. Zou, Y. Zhou, S. Chaudhari, A. Kalyan, T. Rajpurohit, A. Deshpande, K. Narasimhan, V. Murahari, Personagym: Evaluating persona agents and llms, arXiv preprint arXiv:2407.18416 (2024).
[20] B. Peng, Z. Wang, H. Gong, C. Lu, Ip-dialog: Evaluating implicit personalization in dialogue systems with synthetic data, arXiv preprint arXiv:2506.02449 (2025).
[21] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, B. Dolan, deltableu: A discriminative metric for generation tasks with intrinsically diverse targets, arXiv preprint arXiv:1506.06863 (2015).
[22] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, Y. Yu, Texygen: A benchmarking platform for text generation models, in: The 41st international ACM SIGIR conference on research & development in information retrieval, 2018, pp. 1097–1100.
[23] S. Liu, S. Sabour, Y. Zheng, P. Ke, X. Zhu, M. Huang, Rethinking and refining the distinct metric, arXiv preprint arXiv:2202.13587 (2022).
[24] E. Montahaei, D. Alihosseini, M. S. Baghshah, Jointly measuring diversity and quality in text generation models, arXiv preprint arXiv:1904.03971 (2019).
[25] A. Madotto, Z. Lin, C.-S. Wu, P. Fung, Personalizing dialogue agents via meta-learning, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 5454–5459.
[26] C. H. Wu, Y. Zheng, X. Mao, M. Huang, Transferable persona-grounded dialogues via grounded minimal edits, arXiv preprint arXiv:2109.07713 (2021).
[27] Y. Tang, B. Wang, M. Fang, D. Zhao, K. Huang, R. He, Y. Hou, Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona, arXiv preprint arXiv:2305.11482 (2023).
[28] T. Sakai, Z. Zeng, Retrieval evaluation measures that agree with users' serp preferences: Traditional, preference-based, and diversity measures, ACM Transactions on Information Systems (TOIS) 39 (2020) 1–35.
[29] D. Jiang, K. W.-T. Leung, L. Yang, W. Ng, Query suggestion with diversification and personalization, Knowledge-Based Systems 89 (2015) 553–568. URL: https://www.sciencedirect.com/science/article/pii/S0950705115003342. doi:https://doi.org/10.1016/j.knosys.2015.09.003.
[30] Y.-P. Chen, N. Nishida, H. Nakayama, Y. Matsumoto, Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations, 2024. URL: https://arxiv.org/abs/2405.17974. arXiv:2405.17974.
[31] Y. Wang, J. Jiang, M. Zhang, C. Li, Y. Liang, Q. Mei, M. Bendersky, Automated evaluation of personalized text generation using large language models, arXiv preprint arXiv:2310.11593 (2023).
[32] Z. Zhao, C. Vania, S. Kayal, N. Khan, S. B. Cohen, E. Yilmaz, Personalens: A benchmark for personalization evaluation in conversational ai assistants, arXiv preprint arXiv:2506.09902 (2025).
[33] E. J. Gerritse, F. Hasibi, A. P. de Vries, Bias in conversational search: The double-edged sword

of the personalized knowledge graph, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, ICTIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 133–136. URL: https://doi.org/10.1145/3409256.3409834. doi:10.1145/3409256.3409834.
[34] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, arXiv preprint arXiv:1510.03055 (2015).
[35] S. Han, B. Kim, B. Chang, Measuring and improving semantic diversity of dialogue generation, arXiv preprint arXiv:2210.05725 (2022).
[36] F. Guo, A. Metallinou, C. Khatri, A. Raju, A. Venkatesh, A. Ram, Topic-based evaluation for conversational bots, arXiv preprint arXiv:1801.03622 (2018).
[37] K. Stasaski, M. A. Hearst, Pragmatically appropriate diversity for dialogue evaluation, arXiv preprint arXiv:2304.02812 (2023).
[38] K. R. Stasaski, Diversity in Dialogue Generation, University of California, Berkeley, 2022.
[39] R. Sakaeda, D. Kawahara, Generate, evaluate, and select: A dialogue system with a response evaluator for diversity-aware response generation, arXiv preprint arXiv:2206.04937 (2022).
[40] T. Sakai, Comparing two binned probability distributions for information access evaluation, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 1073–1076.
[41] Z. Abbasiantaeb, S. Lupart, L. Azzopardi, J. Dalton, M. Aliannejadi, Conversational gold: Evaluating personalized conversational search system using gold nuggets, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 3455–3465.