DDASR: Deep Diverse API Sequence Recommendation

SIYU NAN and JIAN WANG, School of Computer Science, Wuhan University, Wuhan, China
NENG ZHANG, School of Computer Science and Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China
DUANTENGCHUAN LI and BING LI, School of Computer Science, Wuhan University, Wuhan, China

Recommending API sequences is crucial in software development, saving developers time and effort. While previous studies primarily focus on accuracy, often recommending popular APIs, they tend to overlook less frequent, or "tail," APIs. This oversight, often a result of limited historical data, consequently diminishes the diversity of recommender systems. In this article, we propose DDASR, a framework for recommending API sequences containing both popular and tail APIs. To accurately capture developer intent, we utilize recent Large Language Models for learning query representations. To gain a better understanding of tail APIs, DDASR clusters tail APIs with similar functionality and replaces them with cluster centers to produce a pseudo ground truth. Moreover, a loss function is defined based on learning-to-rank to achieve an equilibrium in accuracy and diversity due to the inherent tradeoff between them. To evaluate DDASR, we conduct extensive experiments on Java and Python open source datasets. Results demonstrate that DDASR significantly achieves the best diversity without sacrificing accuracy. Compared to seven state-of-the-art approaches, DDASR improves accuracy metrics BLEU, ROUGE, MAP, and NDCG and diversity metric coverage by 108.28%, 67.30%, 88.59%, and 45.83%, respectively, on the Java dataset, as well as 9.83%, 2.45%, 8.06%, and 8.03%, respectively, on the Python dataset.

CCS Concepts: • **Software and its engineering → API languages**; **Automatic programming**;

Additional Key Words and Phrases: API Sequence Recommendation, Long-tail Distribution, Clustering, Diversity

ACM Reference format:
Siyu Nan, Jian Wang, Neng Zhang, Duantengchuan Li, and Bing Li. 2025. DDASR: Deep Diverse API Sequence Recommendation. *ACM Trans. Softw. Eng. Methodol.* 34, 6, Article 162 (July 2025), 39 pages.
https://doi.org/10.1145/3712188

This work is supported by the National Key Research and Development Program of China (No. 2022YFF0902701), the National Natural Science Foundation of China (No. 62032016), and the Shenzhen Science and Technology Program (No. CJGJZD20230724091659002).

Authors' Contact Information: Siyu Nan, School of Computer Science, Wuhan University, Wuhan, China; e-mail: siyunan@whu.edu.cn; Jian Wang (corresponding author), School of Computer Science, Wuhan University, Wuhan, China; e-mail: jianwang@whu.edu.cn; Neng Zhang, School of Computer Science and Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China; e-mail: nengzhang@ccnu.edu.cn; Duantengchuan Li, School of Computer Science, Wuhan University, Wuhan, China; e-mail: dtclee1222@whu.edu.cn; Bing Li (corresponding author), School of Computer Science, Wuhan University, Wuhan, China; e-mail: bingli@whu.edu.cn.

ACM 1557-7392/2025/7-ART162
https://doi.org/10.1145/3712188

1 Introduction

An API is a software interface that encapsulates underlying functionalities. Reusing APIs can significantly enhance development efficiency and reduce associated costs [49, 52, 66]. However, with the proliferation of APIs, developers often find it challenging to familiarize themselves with all APIs in the libraries. A survey by Ko et al. [33] indicates that selecting the appropriate API poses a significant learning barrier in user programming systems. Moreover, fulfilling user requirements often necessitates more than just a single API. Developers typically need to look up API sequences to solve tasks. For instance, to address a query like "*Calculates covariance matrix needed assumes an interpolation smoothing for now linear interpolation only diagonal propagates error correctly*," as shown in Figure 1, a sequence of six APIs, including "*scipy.searchsorted*," "*scipy.sqrt*," "*scipy.zeros*," "*scipy.diag_indices_from*," "*scipy.matrix*," and "*scipy.roll*," needs to be invoked.

It is a challenging problem to find appropriate API sequences from the vast array of APIs according to developer requirements [61]. Recently, several approaches have been proposed to recommend APIs for developers. These approaches fall into two major categories: information retrieval-based approaches [26, 59, 72, 73, 80] that search for the most relevant solutions from the historical question repository, and deep learning-based approaches [14, 21, 46] that adopt the **Sequence-to-Sequence (Seq2Seq)** model to recommend API sequences generatively. Gu et al. [21] adopted a **Recurrent Neural Network (RNN)** encoder-decoder and Elnaggar et al. [14] adopted a Transformer encoder-decoder to obtain the results of the API sequence. Martin and Guo [46] applied CodeBERT [16] for the task due to the improved performance of **Large Language Models (LLMs)**. However, existing API recommendation approaches usually recommend popular APIs, neglecting less frequently used ones. Information retrieval-based API recommendation approaches [26, 72] usually match new queries with historical queries in the historical database based on similarity, taking the most similar historical query's ground truth as the solution for the new query. Due to the sparse presence of queries whose ground truths contain APIs with low individual frequencies in the historical database, retrieval-based approaches have limitations in recommending APIs with low individual frequencies [62, 76]. Deep learning-based API recommendation approaches [14, 21, 46] often remove infrequently appearing words from the vocabulary or treat them as <*UNK*> tags, making it challenging to recommend infrequently occurring APIs. In the example in Figure 1, information retrieval-based API recommendation approaches like BIKER [26] and DGAS [72], as well as deep learning-based methods such as DeepAPI [21], CodeBERT [46], and CodeTrans [14], all fail to hit tail APIs. DeepAPI [21] generates the <*UNK*> tag, which holds no practical value for developers.

The long-tail effect [4, 54] occurs when APIs with low individual frequencies collectively constitute a substantial portion. As shown in Figure 2, the long-tail distribution that is objectively present in API libraries can be observed in two open source datasets: a Java dataset [21] and a Python dataset [46]. In the two datasets, APIs occurring twice or less are categorized as tail APIs, whereas, the rest are non-tail APIs. This distinction aligns with the previous deep learning-based approaches [14, 21, 46]. There is a similar distribution pattern in Figure 2(a) and (b). Tail APIs account for a substantial proportion of APIs used in software engineering, representing more than 65% in the open source Java dataset and over 40% in the Python dataset. Zhong and Mei [85] observed that fewer than 10% of APIs are frequently invoked, based on an analysis of over 2 million lines of code in J2SE. Similarly, Ma et al. [45] found that only about 20% of J2SE's APIs are commonly used after examining 39 Java projects. Frequently occurring non-tail APIs are clustered in a small portion at the head, while tail APIs, despite their low individual occurrence frequency, cover a large scale in quantity. The high proportion of tail APIs indicates the developers' need for specific functionalities, underscoring the importance of recommending these APIs in daily programming tasks [11]. From a cognitive perspective, developers are generally more familiar with commonly used APIs

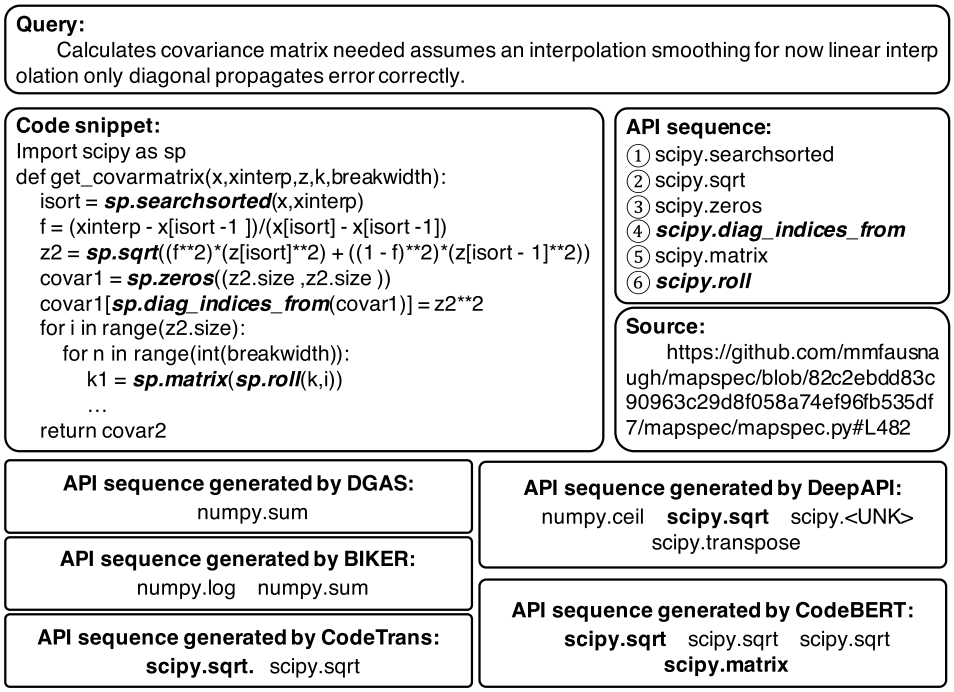

Fig. 1. An example query that needs to be addressed using an API sequence in Java.

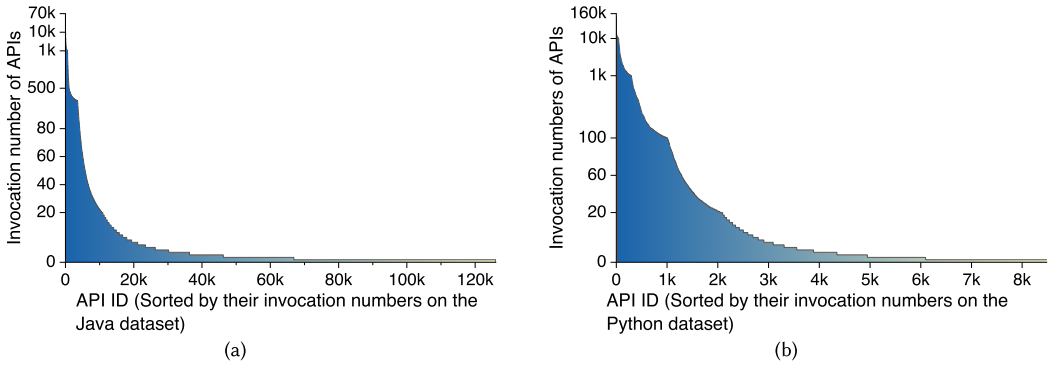

Fig. 2. Long-tail distribution of APIs in the datasets. (a) Distribution in the Java dataset, and (b) distribution in the Python dataset.

(i.e., non-tail APIs) [61], making tail APIs more challenging to learn and use due to the scarcity of existing code samples [85]. This emphasizes the necessity of supporting developers in efficiently utilizing less common, yet essential, APIs. Despite their infrequent appearance, tail APIs may be crucial for developers to solve specific tasks [33, 85]. For instance, as shown in Figure 1, two SciPy APIs, "*scipy.diag_indices_from*" and "*scipy.roll*," are tail APIs that play important roles in computing the covariance matrix, which accounts for errors in linearly interpolated data. "*scipy.diag_indices_from*" is used to access and set the diagonal elements of a square array, while "*scipy.roll*" shifts the elements of an array circularly around the specified axis. These functions are indeed key to calculating a covariance matrix that accurately reflects the propagation of errors through a linear interpolation

process. Additionally, from the perspective of API designers, even though tail APIs are seldom invoked, they are often essential for other APIs and have the potential to become more popular in the future [85]. Hence, learning the representation of tail APIs is essential for API recommendation. However, this is challenging due to the sparse historical usage data of tail APIs [55].

Increasing the aggregated diversity of API sequence recommendation systems, which refers to the variety of different APIs included in the result set generated by the recommendation system [1, 2, 4, 32, 76], can alleviate dependence on popular APIs and enhance the utilization of tail APIs [61, 76]. This, in turn, contributes to the healthy development of the software ecosystem [85]. In addition to expanding the variety of APIs, the balance between the accuracy and diversity of the recommender system should also be considered. As a result, both non-tail and tail APIs must be represented in the recommendation results. However, accuracy and diversity are contradictory metrics, and balancing them becomes another challenge.

To address these issues, we propose a **Deep Diverse API Sequence Recommendation (DDASR)** framework, which can recommend both non-tail and tail APIs and increase diversity without sacrificing accuracy. To better capture developer requirements, we leverage the recent advancements in LLMs. Due to the lack of historical data for tail APIs, their features cannot be adequately acquired. To alleviate the sparsity, we cluster tail APIs with a similarity matrix and substitute them with cluster centers. This strategy helps us establish a pseudo ground truth that includes both non-tail and tail APIs. The similarity matrix is built by calculating the similarity through the description and name of tail APIs. In addition, given that accuracy and diversity are two contradicting indicators, a loss function based on the **Learning-to-Rank (LTR)** technique is defined to optimize the ranking of recommendations while maintaining a balance between accuracy and diversity.

To demonstrate the effectiveness of DDASR, five state-of-the-art API recommendation approaches, i.e., DeepAPI [21], BIKER [26], DGAS [72], CodeBERT [46], and CodeTrans [14], as well as GPT-3.5 and GPT-4, are selected as baselines, with **Bilingual Evaluation Understudy (BLEU)**, **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)**, **Mean Average Precision (MAP)**, **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** and **Normalized Discounted Cumulative Gain (NDCG)** as accuracy evaluation metrics, and coverage as the diversity evaluation metric. Two open source datasets, a Java [21] dataset and a Python [46] dataset, as well as a diverse Java dataset derived from the Java dataset, are employed for evaluation. We evaluate the performance of DDASR utilizing three architectures: RNN encoder-decoder, Transformer encoder-decoder, and LLM encoder-decoder. The RNN and Transformer encoder-decoder architectures aim to reproduce the approaches proposed by Gu et al. [21] and Elnaggar et al. [14]. For the LLM encoder-decoder model, we use the recent five LLMs, such as CodeBERT [16] and GraphCodeBERT [23]. Overall performance is the best when using CodeBERT. In terms of accuracy metrics BLEU, ROUGE, MAP, and NDCG, DDSAR outperforms the baselines by an increase of 63.90%, 34.88%, 46.70%, and 128.60%, respectively, on the original Java dataset, 108.28%, 67.30%, 88.59%, and 207.37%, respectively, on the diverse Java dataset, and 9.83%, 2.45%, 8.06%, 3.26%, respectively, on the Python dataset. Regarding the diversity metric coverage, DDSAR surpasses the baselines by an increase of 45.83% on the diverse Java dataset and 8.03% on the Python dataset. The results indicate that DDASR can significantly increase diversity in recommended API sequences without compromising accuracy. The replication package of DDASR is released at GitHub.[1]

The main contributions of our work are as follows:

—We propose DDASR, a novel framework for recommending API sequences, striking a balance between accuracy and diversity.

[1]https://github.com/WHU-AISE/DDASR

—We focus on the long-tail distribution in API sequence recommendation. We cluster tail APIs with functional similarity and construct a pseudo ground truth by replacing tail APIs in the original ground truth with cluster centers, effectively mitigating the sparsity of historical usage data for tail APIs.
—We conduct extensive experiments to evaluate DDASR using open source Java and Python datasets. A diverse Java dataset is also constructed for further evaluation. The experimental results show that our approach achieves significant diversity without reducing accuracy.

The rest of the article is organized as follows. We present related work in Section 2. We describe the technical details of DDASR in Section 3. We describe our experimental setup and the evaluation results in Section 4. We discuss threats to validity in Section 5. Finally, we conclude the article and put forward future work in Section 6.

2 Related Work

Recommender systems have been extensively studied and applied in software engineering to aid developers in resolving issues of development requirements [20, 61]. This section mainly discusses the literature on recommender systems closely related to this article, including API recommendation, sequence recommendation, and diverse recommendation.

2.1 API Recommendation

API recommendations generally start from developers' requirement queries. Recent research in API recommendation primarily relies on two frameworks: information retrieval and deep learning.

Information Retrieval-Based Approaches. API recommendation approaches based on information retrieval typically perform similarity calculations from the historical library to match existing solutions. Zhong et al. [86] proposed MAPO to mine API usage patterns with sequential rules between APIs. McMillan et al. [47] proposed Portfolio, a tool for finding relevant functions based on PageRank from an extensive archive of C/C++ source code. Chan et al. [8] improved Portfolio using the API graph search algorithm. Raghothaman et al. [58] presented SWIM, an approach designed to find APIs related to the query from user click data and then locate API sequences by extracting and matching structured API sequences from GitHub. RACK [59] recommends APIs by exploiting keyword-API associations from Stack Overflow. Zhang et al. [80] proposed RASH to recommend APIs based on API specifications and the resolution of historical questions. Huang et al. [26] proposed BIKER, which combines historical questions and answers in Stack Overflow with API descriptions. Wei et al. [73] proposed CLEAR, which utilizes BERT sentence embeddings and contrastive learning to capture the semantic information. DGAS [72] utilizes documentation-guided cross-modal attention and documentation-guided cross-modal matching to bridge the gap between text and API. Information retrieval-based API recommendation systems typically operate by matching the current query with historical queries in the database based on similarity, considering the API sequence from the most similar historical query as the solution for the current query. However, due to the sparsity of data in the historical database for queries that tail APIs can resolve, these systems struggle to capture the relevance between queries solvable by tail APIs and the current new query [62, 76]. This limitation makes it difficult for information retrieval-based systems to recommend tail APIs effectively.

Deep Learning-Based Approaches. Deep learning-based generative API recommendation may produce creative results [37]. Niu et al. [50] extracted multiple features and employed LTR to recommend APIs and code examples. DeepAPI [21] builds an open source dataset and customizes the RNN encoder-decoder neural language model for API sequence recommendation. Martin and Guo [46]

developed a Transformer-based neural architecture with CodeBERT [16], an encoder-only pre-trained model, for API sequence learning. CodeTrans [14] uses both supervised and self-supervised tasks to build a language model, which has been applied to API sequence recommendation with the Transformer encoder-decoder architecture. Unfortunately, models such as DeepAPI and CodeTrans often split a complete API into multiple word or symbol fragments, potentially resulting in incorrect API recommendations.

2.2 Sequence Recommendation

A solution to a query involving development requirements often consists of a series of APIs. Classic sequence recommender systems adopt a Markov chain [60] to mine frequent patterns in sequence data. With the development of neural networks, RNN and its variants, such as **Gated Recurrent Unit (GRU)** [24], **Long Short Term Memory (LSTM)** [75], and hierarchical RNN [56], are recognized as beneficial for sequence recommendation. A few works [65, 79] applied convolutional neural networks to sequence recommendation. By adapting attention and self-attention mechanisms, researchers have recently developed many approaches, including NARM [39], SASRec [31], and ATTRec [82]. Recently, researchers have begun exploring the use of LLMs for sequence recommendation tasks [17]. LLMs can be adapted to user data collection, feature engineering, and feature encoder [42]. When applying LLMs, it is important to consider whether to fine-tune them during the training phase. For instance, models like TransRec [18], UNBERT [81], and LMRecSys [83], benefit from fine-tuning during the training phase. In contrast, models such as ChatGPT [13, 25, 43, 64] are designed to operate without the necessity for task-specific fine-tuning during the training phase.

2.3 Diverse Recommendation

Typical approaches for enhancing recommendation diversity can be classified into three categories: pre-processing, in-processing, and post-processing.

Pre-Processing Approaches intervene in the recommendation system before the model training. DCF [12] regards the diverse recommendation as an end-to-end supervised learning task and constructs ground truth labels to explicitly idealize the optimization target. ART [36] leverages a strategy of pre-defining user types to enhance the diversity of recommendations. DGCN [84] includes two pre-defined sampling strategies for diversified recommendations with Graph Convolutional Networks. In addition, the studies [53, 78] utilized long-tail items directly for recommendations. Kim et al. [32] clustered long-tail items to predict a ranked list containing general and diverse items in the next item recommendation.

In-Processing Approaches are applied during the model training process. Wasilewski and Hurley [71] explored the use of regularisation to enhance the diversity of recommendations. Li et al. [40] utilized factorized category features based on matrix factorization. Zhou et al. [87] adapted the Simpson's Diversity Index and considered the evenness of the number of the items' classes.

Post-Processing Approaches re-rank recommended items based on certain diversity metrics. MMR [7] uses a model that treats relevance and diversity as independent metrics. Ziegler et al. [88] presented topic diversification to balance and diversify recommendation results. Adomavicius and Kwon [1] eliminated variations in popularity by assigning different weights to items. Jiang et al. [29] explicitly modeled sub-topics to retrieve diverse results. DDP [10, 74] employs a unified model to assess differences among items in a set-wide way.

3 Approach

Figure 3 shows the overall framework of DDASR, which consists of four main components:

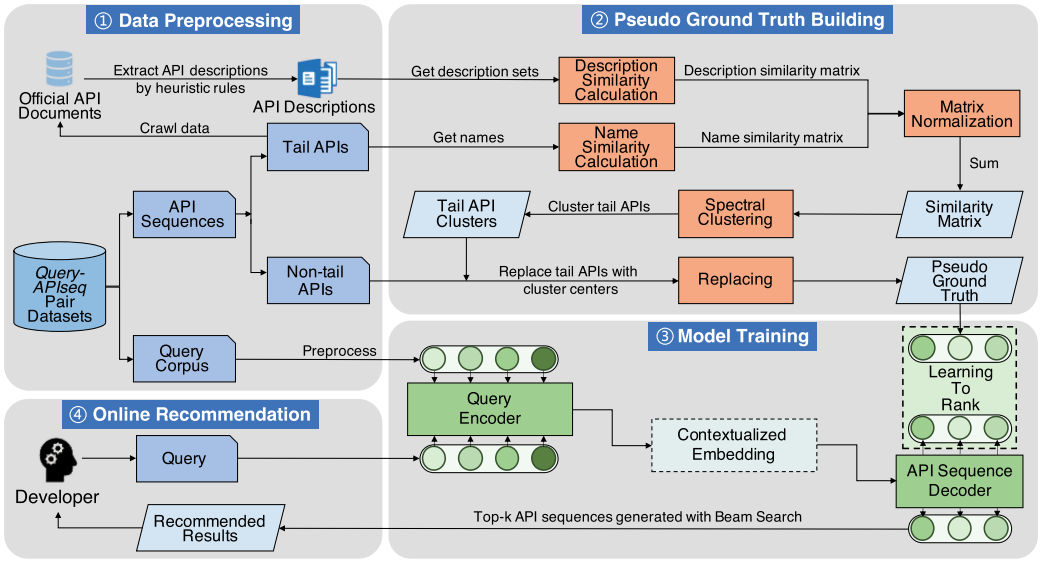

Fig. 3. Overall framework of DDASR.

(1) *Data Pre-Processing* (cf. Section 3.1). We obtain *query-APIseq* pairs, which contain queries and corresponding API sequences. The non-tail and tail APIs are determined by figuring out how frequently they occur. Moreover, we crawl descriptions of tail APIs from the official API documents.
(2) *Pseudo Ground Truth Building* (cf. Section 3.2). We calculate the functional similarity among tail APIs with their descriptions and names. A pseudo ground truth, which contains both non-tail and tail APIs, is established by clustering tail APIs and substituting them with cluster centers.
(3) *Model Training* (cf. Section 3.3). Our model applies two techniques: Seq2Seq (cf. Section 3.3.1) and LTR (cf. Section 3.3.2). The Seq2Seq model is used to translate a given natural language query to a ranked list of possible API sequences. LTR calculates a loss function and ranks these API sequences. We integrate LTR within the Seq2Seq architecture to balance accuracy and diversity in the recommended API sequences. Lastly, Beam Search [34] is utilized to suggest relevant API sequences for developer queries.
(4) *Online Recommendation*. When developers input a query, DDASR can recommend an appropriate API sequence solution.

3.1 Data Pre-Processing

As shown in Figure 4, for query pre-processing, we use the NLTK package [6] to remove stop words and extract the backbone of query terms. In previous generation-based studies [14, 21, 46], an API is typically divided into multiple fragments, where each fragment is a symbol or word representing the class or method name. For example, "*scipy.searchsorted*," an API in Python programming language, will be split into three fragments: "*scipy*," "." and "*searchsorted*," as shown in Figure 4(a). Such division in pre-processing lengthens the API sequence, complicating the modeling of the API calling relationship and dramatically increasing computational costs. Furthermore, this approach often results in mismatched API fragments when generating recommendations for developers. As shown in Figure 1, the deep learning-based method DeepAPI [21] recommends the API fragments

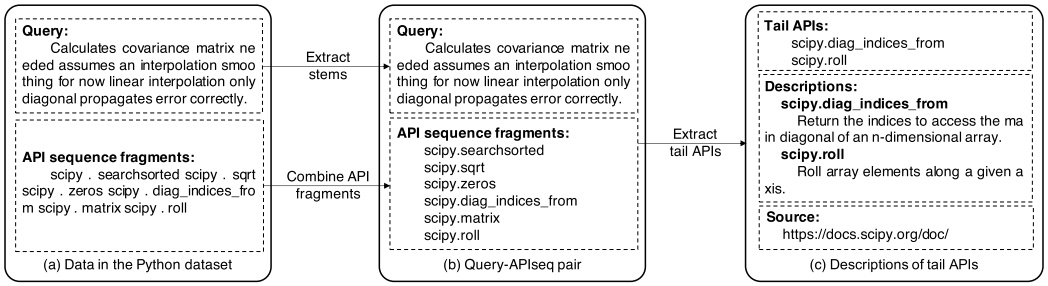

Fig. 4. An example for data pre-processing phase.

"*scipy*," "." and "*transpose*," which are combined to form the complete API name "*scipy.transpose*." However, upon analysis, we find that this is an incorrect API, highlighting issues in combining class and method names.

Inspired by the word embedding technique [5], we recombine these fragments into complete APIs, allowing the model to better learn the calling relationship between APIs and reducing the computation of decoding APIs. After pre-processing the queries and API fragments, the corpus containing *query-APIseq* pairs for Java and Python are obtained, as shown in Figure 4(b). Subsequently, we conduct a frequency analysis of API occurrences and distinguish between non-tail and tail APIs.

For tail APIs, we mine their descriptions to facilitate functional similarity calculations, as illustrated in Figure 4(c). To achieve this, we download both official API documentation and third-party documents. Then we parse the HTML file of each API class to extract their descriptions. Generally, APIs with inheritance relationships tend to have similar functions. Therefore, we also mine descriptions of APIs within these inheritance relationships as a supplementary measure.

3.2 Pseudo Ground Truth Building

Traditional API sequence recommendation approaches often overlook tail APIs, either by ignoring them or labeling them as <*UNK*> tags, which leads to these APIs being under-recommended. Noting that APIs with similar functions typically share analogous functionalities, we cluster tail APIs with similar functionality and then substitute them with cluster centers to build a pseudo ground truth. This pseudo ground truth includes both non-tail and tail APIs, ensuring comprehensive recommendations. In this context, we use TA and NA to represent tail APIs and non-tail APIs, respectively.

3.2.1 Similarity Calculation. As a primary indicator, the name of an API typically reflects its function, while the official API documentation describes its functionalities. In this section, we calculate the functional similarity among tail APIs using both their documentation descriptions and names [35]. A functional similarity matrix is ultimately constructed, containing the functional similarity between each pair of tail APIs.

Similarity Calculation of API Documentation Descriptions. We observe that official documentation contains numerous inheritance associations among APIs. To augment the descriptions of child APIs, we append the official descriptions of their parent APIs. Designing APIs with a preference for shallow inheritance hierarchies is considered a good design principle and is widely accepted and recommended [19]. Therefore, we only consider API descriptions that involve inheritance relationships of parent nodes. Let $\mathbb{D}_i = \{D_{i,s}, D_{i,p}\}$ represent the description set of tail API TA_i, where $D_{i,s}$ is the description of TA_i itself, and $D_{i,p}$ denotes the description inherited from the parent node of TA_i.

We use BM25, an algorithm for evaluating the similarity between search statements and documents in the corpus, to calculate the description similarity between tail APIs. To assess the text-similarity between two descriptions, D_1 and D_2, we approach it in a bifurcated manner: $Sim(D_1 \rightarrow D_2)$ and $Sim(D_2 \rightarrow D_1)$. These are calculated using the BM25 algorithm, yielding two asymmetric scores. This approach recognizes that the impact of a morpheme can vary across different sentences, necessitating a nuanced evaluation of similarity. Then the harmonic mean is taken as the similarity score $Sim(D_1, D_2)$ between the two asymmetric scores:

$$Sim\,(D_1, D_2) = \frac{2 \times Sim\,(D_1 \rightarrow D_2) \times Sim\,(D_2 \rightarrow D_1)}{Sim\,(D_1 \rightarrow D_2) + Sim\,(D_2 \rightarrow D_1)}. \tag{1}$$

Due to the inheritance relationship, each tail API description set $\mathbb{D}$ includes both its description and the descriptions of its parents. However, the description of its parents cannot fully reflect the functionalities of a tail API. Therefore, we set up a semaphore DI as the weight to distinguish the impact of descriptions from different sources. According to [67], we set DI to 2 if the description similarity is calculated based on the tail API's description; otherwise, DI is set to 1. After iterating through all descriptions in $\mathbb{D}$, we take the maximum similarity as the description similarity $simD$ of two tail APIs:

$$simD(TA_i, TA_j) = max\{DI \times Sim(D_{i,k}, D_{j,m}) | D_{i,k} \in \mathbb{D}_i \; and \; D_{j,m} \in \mathbb{D}_j\}. \tag{2}$$

Similarity Calculation of API Names. We use the Levenshtein distance [38], a minimum edit distance algorithm, to calculate the similarity between the names of tail APIs. The Levenshtein distance is computed as the minimum number of three single-character edit operations, including deletion, insertion, and substitution (matching or mismatching), for converting one API name to another:

$$Lev_{N_i,N_j}(k,m) = \begin{cases} max(k,m), & min(k,m) = 0 \\ min\begin{pmatrix} Lev_{N_i,N_j}(k-1,m)+1, \\ Lev_{N_i,N_j}(k,m-1)+1, \\ Lev_{N_i,N_j}(k-1,m-1)+1_{N_i[k] \neq N_j[m]} \end{pmatrix}, & otherwise \end{cases}, \tag{3}$$

where N_i and N_j represent the names of tail APIs TA_i and TA_j, respectively. The indices k and m correspond to the positions within N_i and N_j, respectively. The function $Lev_{N_i,N_j}(k,m)$ calculates the distance between the first k characters of N_i and the first m characters of N_j. In the minimum operation, the first term accounts for deletion implying the removal of a character from N_i to match N_j. The second term accounts for insertion, denoting the addition of a character to N_i. The final term represents the cost of substitution, which is determined by the indicator function $1_{N_i[k] \neq N_j[m]}$. The function yields 1 if $N_i[k] \neq N_j[m]$, indicating a mismatch, and 0 if they match.

The name similarity, $simN$, is calculated based on the Levenshtein distance:

$$simN(TA_i, TA_j) = 1 - \frac{Lev_{N_i,N_j}(|N_i|, |N_j|)}{max(|N_i|, |N_j|)}, \tag{4}$$

where $|N_i|$ and $|N_j|$ represent the length of N_i and N_j, respectively. $Lev_{N_i,N_j}(|N_i|, |N_j|)$ equals 0 when $N_i = N_j$.

Functional Similarity Matrix. Due to the different calculation methods of description similarity $SimD$ and name similarity $SimN$, they exhibit a significant difference in scale. Therefore, we employ *Min-Max Normalization* to standardize their scales. When the normalized similarity measures $simD$ and $simN$ exceed respective thresholds δ_1 and δ_2, the APIs in question are considered to have related descriptions or names. Otherwise, their similarity is deemed irrelevant and is effectively set to zero. The values of α and β will be discussed in Section 4.6.4, where α and β represent the

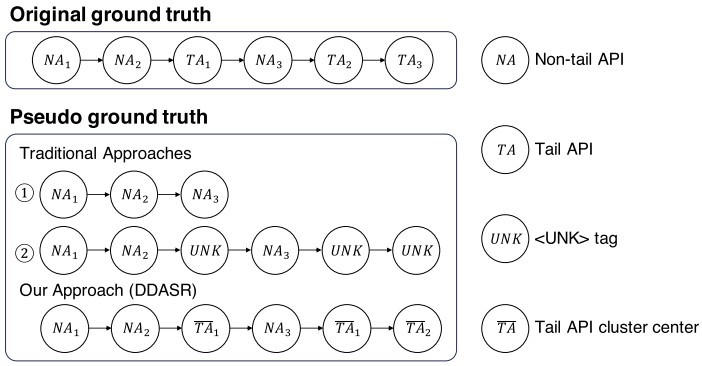

Fig. 5. The original ground truth and the pseudo ground truth of API sequence recommendation.

weights assigned to $simD$ and $simN$, respectively, under the condition that $\alpha + \beta = 1$. The final similarity score, Sim, is calculated by adding $simD$ and $simN$ after normalizing them with weights:

$$Sim(TA_i, TA_j) = \alpha \times max(\frac{simD(TA_i, TA_j)}{simD_{max} - simD_{min}} - \delta_1, 0) + \beta \times max(\frac{simN(TA_i, TA_j)}{simN_{max} - simN_{min}} - \delta_2, 0). \tag{5}$$

We calculate the similarity of all tail APIs in pairs to obtain an $n \times n$ symmetric similarity score matrix **SM**.

3.2.2 Clustering and Replacing. To improve the representation of tail APIs, we build a pseudo ground truth by clustering the tail APIs based on the functional similarity matrix and replacing the tail APIs with cluster centers.

For a given query Q in the original ground truth, the answer is a sequence that consists of a series of APIs, such as $< NA_1, NA_2, TA_1, NA_3, TA_2, TA_3 >$. Due to the sparse historical usage data of tail APIs, directly using the original ground truth, which contains a large number of tail APIs, poses significant challenges to accuracy [28, 44, 63]. Traditional deep learning-based API sequence recommendation approaches [14, 21, 46] treat the sequence as $< NA_1, NA_2, NA_3 >$ or $< NA_1, NA_2, < UNK >, NA_3, < UNK >, < UNK >>$ since they delete tail APIs, i.e., TA_1, TA_2, and TA_3, or use $<UNK>$ tags to replace them in the pre-processing phase as shown in Figure 5. In addition, we conduct experiments using the original ground truth and the pseudo ground truth (where tail APIs are treated as $<UNK>$ tags) on two deep learning-based methods, DeepAPI and CodeBERT. The experimental results shown in Table 1 demonstrate that the accuracy of both models dramatically decreases when directly using the original ground truth compared to the pseudo ground truth. This indicates that the processing of the pseudo ground truth is effective in improving the accuracy of deep learning-based models. However, in existing methods [14, 21, 46], tail APIs are treated as $<UNK>$ tags in constructing the pseudo ground truth, which prevents the recommendation of these important APIs to developers. Thus, we apply spectral clustering [68] to group tail APIs based on functional similarity, effectively representing similar tail APIs with their cluster centers.

We treat the similarity matrix **SM** of tail APIs as an adjacency matrix and use it to calculate a diagonal symmetric matrix with dimension $n \times n$, denoted as $\mathbf{D}^{iag}$:

$$\mathbf{D}^{iag}{}_{ij} = \begin{cases} \sum_k \mathbf{SM}_{ik}, & i = j,\ 0 \le k < n \\ 0, & i \neq j \end{cases}. \tag{6}$$

Table 1. Evaluation Results of DeepAPI and CodeBERT with Different Ground Truths on the Java and Python Datasets

Models	Type of Ground Truth	Java Dataset	Python Dataset
		$BLEU_O$	
DeepAPI	OGT	20.18	39.61
	PGT(<UNK>)	28.65	47.54
CodeBERT	OGT	42.11	47.93
	PGT(<UNK>)	50.86	52.40

"OGT" represents the original ground truth and "PGT(<UNK>)" represents the pseudo ground truth built by replacing tail APIs with <UNK> tags.

Algorithm 1: Generate Sample Points for Spectral Clustering

Input: **SM**: Similarity matrix of tail APIs. c: number of clusters.
Output: $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$: Sample points.
Begin
1: /* Compute the diagonal matrix. */
2: $\mathbf{D}^{iag} \leftarrow diag(\sum_k \mathbf{SM}_{0k}, \sum_k \mathbf{SM}_{0k}, \ldots, \sum_k \mathbf{SM}_{0k})$
3: /*Compute the normalized Laplace matrix. */
4: $\mathbf{L}_{rsym} \leftarrow \mathbf{D}^{iag\,-1/2}(\mathbf{D}^{iag} - \mathbf{SM})\mathbf{D}^{iag\,1/2}$
5: /*Compute the first c eigenvectors $u_1, u_2, \ldots, u_c$ of $\mathbf{L}_{rsym}$. */
6: $\mathbf{U} \leftarrow [u_1, u_2, \ldots, u_c] \in \mathbf{R}^{n \times c}$
7: /*Compute matrix $\mathbf{T}$ from matrix $\mathbf{U}$ by individually normalizing each row with L_2-norm. */
8: $\mathbf{T} \leftarrow [t_{ij} = u_{ij} / \sqrt{\sum_{m=1}^{n} u_{im}^2}] \in \mathbf{R}^{n \times c}$
9: /*Define $y_i \in \mathbf{R}^c$ as the vector associated with the ith row of matrix $\mathbf{T}$. */
10: $\mathbf{Y} \leftarrow \{y_i = row_i(\mathbf{T})\}_i^n$
11: **return** $\mathbf{Y}$

Based on Equation (6), we define a normalized symmetric Laplace matrix $\mathbf{L}_{rsym}$:

$$\mathbf{L}_{rsym} = \mathbf{D}^{iag\,-\frac{1}{2}}(\mathbf{D}^{iag} - \mathbf{SM})\mathbf{D}^{iag\,\frac{1}{2}}. \tag{7}$$

Algorithm 1 generates sample points based on tail APIs and their similarity matrix **SM** for clustering. We begin by computing $\mathbf{D}^{iag}$ and $\mathbf{L}_{rsym}$ (Lines 1–4). Then we calculate the eigenvalues of $\mathbf{L}_{rsym}$ and select the first c eigenvalues after sorting in ascending order. After calculating the eigenvectors of eigenvalues, we take them as column vectors to form the matrix $\mathbf{U}$ (Lines 5–6). Finally, the vectors in each row of $\mathbf{U}$ are converted to unit vectors to form new sample points $\mathbf{Y}$ (Lines 7–10).

After that, we use the k-means algorithm to assign tail APIs with similar functionality into a cluster $\overline{\mathbf{TA}} = \{\overline{\mathbf{TA}}_1, \overline{\mathbf{TA}}_2, \ldots, \overline{\mathbf{TA}}_j, \ldots \overline{\mathbf{TA}}_c\}$ based on sample points Y, where c represents the number of cluster centers. For the cluster $\overline{\mathbf{TA}}_j$, $\overline{TA}_j$ is the cluster center. Recommending cluster centers can expose some hidden tail APIs, increase the diversity of recommendations, and offer additional perspectives and possibilities, thereby inspiring developers' creativity due to the functional similarity of the exposed and hidden tail APIs in the clusters [9]. Therefore, we substitute tail APIs with cluster centers. For an API sequence represented as $< NA_1, NA_2, TA_1, NA_3, TA_2, TA_3 >$, we restructure it to $< NA_1, NA_2, \overline{TA}_1, NA_3, \overline{TA}_1, \overline{TA}_2 >$ to build a pseudo ground truth, as illustrated

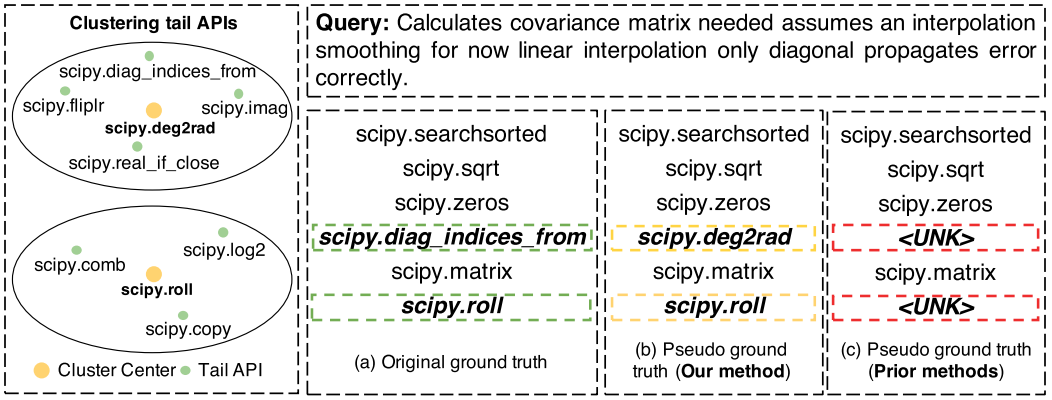

Fig. 6. An example of building the pseudo ground truth. (a) The original ground truth, (b) the pseudo ground truth built by clustering tail APIs and replacing them with cluster centers (as done in our method), and (c) the pseudo ground truth built by replacing tail APIs with <*UNK*> for the query (as done in prior methods).

in Figure 5. This restructuring is based on the categorization of TA_1 and TA_2 as belonging to the same cluster $\overline{\mathbf{TA}}_1$, while TA_3 is classified under a different cluster $\overline{\mathbf{TA}}_2$. Figure 6 shows an example of building the pseudo ground truth. There are two clusters of tail APIs as shown in Figure 6. For the query, Figure 6(a) is the original ground truth, Figure 6(b) is the pseudo ground truth built by clustering tail APIs and relocating with the cluster centers, and Figure 6(c) is the pseudo ground truth constructed by replacing tail APIs with <*UNK*> tags. Moreover, the number of occurrences of each cluster is the sum of the number of occurrences of tail APIs belonging to the cluster. However, the clusters serve as substitutes for the tail APIs in the training process. Each tail API cluster should not appear more frequently than the non-tail APIs. Thus, we set the number of clusters c such that the average number of tail APIs in clusters approximates that of non-tail APIs, and ensure that the item count in each cluster does not substantially exceed that of non-tail APIs.

3.3 Model Training

3.3.1 Seq2Seq Model. DDASR applies Seq2Seq, an encoder-decoder model, to generate API sequences. For a given query, the encoder can produce a contextualized embedding vector, which is used by the decoder to produce an API sequence. The LLM encoder-decoder architecture is used in DDASR, with LLMs serving as the encoder to capture developer requirements and a six-layer Transformer acting as the decoder.

3.3.2 LTR Loss Function. To achieve a balance between accuracy and diversity, we adopt ListMLE [77], a listwise LTR method, as our loss function. By utilizing ListMLE to model the rankings of all API sequences generated, we learn the order of ranking to balance the effect of tail API clusters on diversity by improving the prediction accuracy of the rankings, thus allowing the model to increase diversity without sacrificing accuracy. For a given query Q and corresponding API sequence A in the pseudo ground truth, we define a sequence probability distribution based on Plackett-Luce [77] for the recommendation results as follows:

$$P(\mathbf{A}|s) = \prod_{i=1}^{m} \frac{\varphi(s_{A^{-1}(i)})}{\sum_{u=i}^{m} \varphi(s_{A^{-1}(u)})}, \tag{8}$$

where $A^{-1}(i)$ represents the APIs sorted to the ith positions in sequence $\mathbf{A} =< A_1, A_2, \cdots, A_m >$. $s_{A^{-1}(i)}$ is a ranking score of item A_i from ranking scores vector $s = f(x)$ and $\varphi()$ is an exponential

mapping function, namely $\varphi(x) = exp(x)$. The better the fit between $\mathbf{A}$ and s, the higher the probability value $P(\mathbf{A}|s)$.

We define our loss function L as the inverse of log-likelihood:

$$L = -logP(\mathbf{A}|f(x)), \tag{9}$$

where L is a convex function so that we can optimize it directly with gradient descent.

Finally, DDASR produces top-k API sequences for each given query by using Beam Search [34], a heuristic search strategy based on the average loss.

4 Evaluation

In this section, we evaluate the proposed DDASR approach. We will study the following four **Research Questions (RQs)**:

- —RQ1: How accurate is DDASR in comparison with the state-of-the-art approaches?
- —RQ2: Does DDASR improve diversity while maintaining accuracy?
- —RQ3: Can DDASR help programmers address tasks more effectively?
- —RQ4: How do the different functional similarity algorithms and the parameter weights in the similarity algorithms affect the performance of DDASR?

4.1 Experiment

All our experiments were performed on a Supermicro GPU server with Intel(R) Xeon(R) E5-2640 v4 x86_64, 2.4 GHz, 2 GPUs (NVIDIA Tesla V100 16 GB), and the CentOS 7.5 Linux operating system. The machine learning models were implemented with the PyTorch library.

4.2 Data

We utilize two open source datasets: one for the Java programming language [21] and the other for the Python programming language [46]. The Java dataset constructed by Gu et al. [21] consists of 7 million annotation-API sequence pairs. These are extracted by mining Java projects on GitHub that have garnered more than one star. We treat the natural language annotations as queries. However, in the original Java dataset, we find about 6 million duplicate records and many errors in the API sequences, as illustrated in Table 2. We eliminate the duplicate records. Subsequently, our examination revealed a total of 25,862 records containing mismatched symbols, with the predominant issue being a disparity in the count of left and right parentheses. To address this imbalance, we primarily rectified the records by adding parentheses where necessary to achieve symmetry. In addition, we find 3,552 records whose class or method names mismatch the official API documentation. We revise them manually by consulting the API documentation. Our developers manually review approximately 20k processed APIs. Finally, we obtain a repository in Java, including about 760k *query-APIseq* pairs and 120k APIs. Additionally, we observe a phenomenon that the test set of the original dataset contains very few *query-APIseq* pairs of tail APIs. To evaluate the diversity in API sequence recommendation, we build a diverse Java dataset with an equal distribution of pairs containing tail APIs in the training and test sets by stratified random sampling. The Python dataset constructed by Martin and Guo [46] consists of about 855k records by collecting Python projects with at least five stars. After a comprehensive investigation, we do not find the issue shown in Table 2. Moreover, the Python dataset contains a balanced record of tail APIs in both the test and training sets and can be used to evaluate diversity without constructing a separate diverse dataset. Finally, we obtain the Python dataset that contains around 855k *query-APIseq* pairs and 8.5k APIs. Table 3 shows the summary of the datasets.

Table 2. Five Data Cleaning Steps on the Original Java Dataset with Corresponding Scenarios, Processing Methods, and Examples

Type	Scenarios	Processing Method	Example	Example after Processing
Duplicate pairs	Duplicate records for the same *query-APIseq* pairs in the dataset.	Delete duplicate records and keep only one.	-	-
Class name errors	Class name is inconsistent with the official API documentation.	Change to standard class name according to the official API documentation.	CDRInputStream_1_0.<init >	CDRInputStream.<init >
Method name errors	Method name is inconsistent with the official API documentation.	For those caused by spelling errors, make the necessary corrections. For method names that do not exist, remove the entire API.	ConcurrentLinkedDeque.succ	-
	Not all actual arguments have been completely removed.	Remove the actual arguments from the method name.	StringBuilder.append("time")	StringBuilder.append
Mismatched symbols	The number of left and right parentheses is unequal.	Complete the mismatched parentheses.	ArrayList<.append	ArrayList<>.append

Table 3. Summary of the Datasets

Datasets		# Pairs	# Pairs Containing Tail APIs	# Non-Tail APIs	# Tail APIs
Java dataset	Training set	7,519,907	63,239	46,245	79,797
	Test set	9,975	175	12,895	227
Diverse Java dataset	Training set	752,919	62,654	46,180	78,521
	Test set	7,600	760	12,198	1,555
Python dataset	Training set	835,069	3,914	4,949	3,534
	Test set	10,000	131	1,914	153

For the descriptions of tail APIs, we mine the API documentation based on the class names of long-tail APIs. When experimenting on Java, we download the Java SE 8 API documentation[2] and third-party documentation, such as Spring Boot,[3] JUnit,[4] and Log4j.[5] When experimenting on Python, we crawl the official documentations of Python 3,[6] Numpy,[7] SciPy,[8] PyTorch,[9] and Pandas[10] to mine the descriptions.

[2]http://www.oracle.com/technetwork/java/javase/
[3]https://docs.spring.io/spring-boot/docs/2.6.12/api/
[4]https://junit.org/junit4/javadoc/latest/
[5]https://logging.apache.org/log4j/2.x/log4j-api/apidocs/
[6]https://docs.python.org/3.8/
[7]https://numpy.org/doc/stable/reference/
[8]https://docs.scipy.org/doc/scipy/reference/
[9]https://pytorch.org/docs/stable/
[10]https://pandas.pydata.org/docs/reference/

4.3 Baselines

We selected seven state-of-the-art approaches as competing models: DeepAPI [21], BIKER [26], DGAS [72], CodeBERT [46], CodeTrans [14], GPT-3.5, and GPT-4. Among them, BIKER and DGAS are information retrieval-based methods, while GPT-3.5 and GPT-4 are generative models capable of producing API sequences based on input prompts. These two LLMs were chosen as baselines due to their superior performance across various natural language processing tasks compared to other open source and closed-source LLMs [51]. The remaining models, DeepAPI, CodeBERT, and CodeTrans, are generative approaches based on deep learning techniques.

- *DeepAPI*[11] [21] adapts an RNN encoder-decoder model to encode a query into a fixed-length context vector and generate API sequences based on the context vector.
- *BIKER*[12] [26] extracts API-related posts of Stack Overflow, treating the titles of the questions as search queries and the APIs mentioned in the accepted answers as the canonical solutions. The tool mainly calculates the similarity of questions and recommends the APIs of similar questions. To adapt BIKER for API sequence recommendation, we make the necessary modifications to its open source code. The detail of modifications is shown on GitHub.[13]
- *DGAS*[14] [72] utilizes a multi-head self-attention network and trains a deep learning network for API sequence search through three phases: the inner-modal attention phase, the documentation-guided attention phase, and the documentation-guided cross-modal attention phase. In the searching phase, DGAS searches for a ranked list of relevant API sequences based on the matching scores calculated between the query and API sequences by the trained network.
- *CodeBERT*[15] is reproduced by Martin and Guo [46] to generate API sequences by fine-tuning CodeBERT [16], a pre-trained model, on the Java and Python datasets.
- *CodeTrans*[16] [14] combines encoder-decoder and Transformer models and explores different training strategies to recommend API sequences. In addition, CodeTrans provides a pre-trained model for the task of API sequence generation on the original Java dataset, which we utilize directly for experimental comparison.
- *GPT-3.5*[17] is a model improved on GPT-3. We use OpenAI's APIs to access its latest version, gpt-3.5-turbo.
- *GPT-4*[18] is a large multi-modal with broader general knowledge and advanced reasoning capabilities. We also use OpenAI's APIs to access its latest version, gpt-4-turbo.
- *Ablation Studies.* Our DDASR comprises three key components: recombining fragments into complete APIs, building the pseudo ground truth, and training the Seq2Seq model with a LTR loss function. Since the three components in our DDASR are in an incremental relationship, where the second component depends on the first and the third component depends on the second, we design two variants of DDASR, DASR (containing only the first component), and DASR+PGT (containing the first two components), for ablation studies. Our DDASR incorporates all three components.

[11] https://github.com/guxd/deepAPI
[12] https://github.com/tkdsheep/BIKER-ASE2018
[13] https://github.com/WHU-AISE/DDASR
[14] https://github.com/helloDGASworld/DGAS_dataset
[15] https://github.com/hapsby/deepAPIRevisited
[16] https://github.com/agemagician/CodeTrans
[17] https://platform.openai.com/docs/models/gpt-3-5-turbo
[18] https://platform.openai.com/docs/models/gpt-4-turbo

– *DASR* is a variant of our DDASR, which is designed to be an ablation study to assess the effect of the first component—merely recombining API fragments. In DASR, API fragments are combined into complete APIs during the construction of the pseudo ground truth, with tail APIs treated as <UNK> tags. The loss function of DASR is Cross Entropy, following the previous work [21, 46].
– *DASR+PGT* is to assess the impact of integrating DASR with the pseudo ground truth component. DASR+PGT constructs pseudo ground truth using the cluster centers of tail APIs. Its loss function is also Cross Entropy.

4.4 Model Selection

We consider three types of encoder-decoder models in our experiment.

— *RNN Encoder-Decoder Model.* We choose the model because it is used to generate API sequences by Gu et al. [21]. A standard Bidirectional LSTM is employed in the encoder, and the decoder is a GRU network.
— *Transformer Encoder-Decoder Model.* The encoder and decoder consist of six layers of Transformer encoder and decoder components. CodeTrans [14] also employs a similar architecture to accomplish tasks.
— *LLM Encoder-Decoder Model.* LLMs currently demonstrate outstanding performance in various natural language processing tasks. They are categorized into three architectures: encoder-only, decoder-only, and encoder-decoder. Decoder-only models have vocabulary limitations and heavily rely on the presence of tokens in their vocabulary for prediction [57]. The APIs in the pseudo ground truth are not included in the vocabulary of decoder-only models, resulting in no matches with any tokens in the model's fixed vocabulary [63]. Additionally, our loss function based on LTR is specifically designed for API sequences where tokens correspond to complete APIs. Therefore, in our approach, we have not considered decoder-only models. For the encoder-only and encoder-decoder LLMs, we only utilize their encoder component as the encoder for our approach. We select five widely available LLMs for code in the HuggingFace API,[19] including CodeBERT,[20] GraphCodeBERT,[21] PLBART,[22] CodeT5,[23] and UniXcoder.[24] CodeBERT [16] is an encoder-only model pre-trained on bimodal data sourced from CodeSearchNet [27]. GraphCodeBERT [23] is also an encoder-only model that incorporates syntax information from code. The other three LLMs are encoder-decoder models. PLBART [3] is pre-trained on a wide range of Java and Python function collections and related natural language text using denoising auto-encoders. CodeT5 [70] can better leverage the code semantics conveyed from the developer-assigned identifiers. UniXcoder [22] comprehensively utilizes the code structure information provided by the Abstract Syntax Tree. The decoder consists of six layers of Transformer decoder.

We use the following hyper-parameter settings. Based on our experiment results shown in our reproducible package on GitHub, we select the optimal hyper-parameters when using RNN encoder-decoder architecture. For the RNN encoder-decoder, we set the number of hidden layers, the number of hidden units, and the dimension of word embedding to 3, 1,000, and 512, respectively.

[19] https://huggingface.co
[20] https://huggingface.co/microsoft/codebert-base
[21] https://huggingface.co/microsoft/graphcodebert-base
[22] https://huggingface.co/uclanlp/plbart-base
[23] https://huggingface.co/Salesforce/codet5-base
[24] https://huggingface.co/microsoft/unixcoder-base

For the Transformer encoder-decoder, both the encoder and decoder have six layers, and the word embedding dimension is 512, which is the same as that used in the RNN encoder-decoder architecture. For the LMM encoder-decoder, we utilize the pre-defined hyper-parameters from the openly available models on HuggingFace. Specifically, the hidden dimensions for the LMMs used as the encoder are 768, and the decoder has six layers. The AdamW algorithm is used to optimize model parameters with the learning rate $\alpha = 2e - 5$. Moreover, we adopt the optimal settings of baselines reported in the literature for fair comparison.

4.5 Metrics

—*BLEU* is used to evaluate how close the candidate API sequences generated by the model are to the reference API sequence in the ground truth. BLEU calculates the *n-gram* hit ratio of the candidate sequence in the reference sequence. The calculation formula of BLEU is as follows:

$$BLEU@K = \frac{1}{K}\sum_{k=1}^{K} BP@k \cdot exp(\sum_{n=1}^{N}\frac{1}{n} log\frac{C'_{ref}n + 1}{C_{can_k}n + 1}), \tag{10}$$

where K represents the top-K API sequences recommended, and N represents the largest number of grams. N is usually set to 4. $C_{can_k}n$ refers to the number of the *n-gram* in the kth recommended API sequence. ref represents the reference API sequence that is the correct answer in the ground truth. $C'_{ref}n$ refers to the number of *n-gram* in the reference API sequence simultaneously. *BP* is a length penalty factor to avoid the effect of too short sequences on the effectiveness of evaluation results. *BP* is defined as

$$BP@k = \begin{cases} 1 & L_{can_k} > L_{ref} \\ e^{1-L_{ref}/L_{can_k}} & L_{can_k} \leq L_{ref} \end{cases}, \tag{11}$$

where L_{ref} refers to the length of the reference API sequence, and L_{can_k} refers to the length of the kth candidate API sequence.

—*ROUGE* is an evaluation method that compares candidate API sequences with the reference API sequence based on recall. We primarily consider the ROUGE-L metric, which focuses on the **Longest Common Subsequence (LCS)** between a candidate API sequence and the reference API sequence. Formally,

$$ROUGE@K = \frac{1}{K}\sum_{k=1}^{K} \frac{2 \cdot \frac{LCS(ref,can_k)}{L_{ref}} \cdot \frac{LCS(ref,can_k)}{L_{can_k}}}{\frac{LCS(ref,can_k)}{L_{ref}} + \frac{LCS(ref,can_k)}{L_{can_k}}}, \tag{12}$$

where $LCS(ref, can_k)$ denotes the length of the longest common subsequence between ref and can_k.

—*MAP* is the mean of the AP score for each query. Formally,

$$MAP@K = \frac{1}{L_{ref}}\sum_{A \in ref}\frac{1}{K}\sum_{k=1}^{K}\frac{num_A(k)}{k}, \tag{13}$$

where $num_A(k)$ represents the number of APIs denoted as A in the top-k candidate API sequence.

—*NDCG* evaluates the ranking quality of candidate API sequences. Formally,

$$NDCG@K = \frac{1}{L_{ref}}\sum_{A \in ref}\frac{DCG_A@K}{IDCG@K}, \tag{14}$$

where $DCG_A@K$ uses graded relevance as a measure of gain, the gain is discounted according to its ranking position. We define it as

$$DCG_A@K = \sum_{k=1}^{K} \frac{2^{r_A^{(k)}} - 1}{log_2(k+1)}, \tag{15}$$

where $r_A^{(k)}$ is a relevance score. $r_A^{(k)}$ is set to 1 if the API appears in the kth candidate sequence, otherwise 0. Since the gain with a high rank is high, the low-ranked score is somewhat compromised according to the order in which the API appears in the candidate API sequences. $IDCG@K$ is equivalent to $max(DCG_A@K)$, which normalizes $DCG_A@K$. Formally,

$$IDCG@K = \sum_{k=1}^{K} \frac{1}{log_2(k+1)}. \tag{16}$$

— *Coverage* [30] is used to assess the diversity of API sequence recommender systems. It is defined as the proportion of all candidate API sequences exposed to APIs in the vocabulary.

$$Coverage = \frac{\#item\ in\ candidate\ API\ sequences}{\#total}. \tag{17}$$

BLEU, ROUGE, MAP, and NDCG are metrics for evaluating the accuracy. However, previous work has typically used the pseudo ground truth as a reference for evaluating, which we refer to as falsified accuracy. This is because, in the pseudo ground truth, tail APIs are replaced with <*UNK*> tags or cluster centers, leading to evaluation bias. We denote the accuracy metrics evaluated using the pseudo ground truth as BLEU_P, ROUGE_P, MAP_P, and NDCG_P. To restore the veritable accuracy, we also utilize the original ground truth as a reference, and in this case, the accuracy metrics are labeled as BLEU_O, ROUGE_O, MAP_O, and NDCG_O. When assessing the falsified accuracy of different approaches, we use the pseudo ground truth constructed by each approach itself as the reference. Conversely, the original ground truth is uniformly used as the reference when assessing the veritable accuracy of any approach. For example, our DDASR constructs the pseudo ground truth using cluster centers of tail APIs, as shown in Figure 6(b). This pseudo ground truth is then used as the reference when evaluating DDASR's falsified accuracy. Similarly, the prior deep learning-based methods construct the pseudo ground truth using <*UNK*> tags, as shown in Figure 6(c), and this is used as the reference when evaluating the falsified accuracy of deep learning-based baselines. Additionally, when assessing the veritable accuracy of all methods, the original ground truth, as shown in Figure 6(a), is used as the same reference. Specifically, the IR-based baselines do not construct the pseudo ground truth. Therefore, when evaluating the falsified accuracy of the IR-based baselines, the original ground truth is regarded as the reference, similar to the evaluation of the veritable accuracy.

Moreover, considering the effectiveness and practicality of the API sequence recommender, the smallest unit of evaluation in our experiments is a complete API, which includes both the class name and method name in its full designation. This approach helps avoid invalid combinations of API classes and method names, providing directly usable and correct APIs that enhance developers' work efficiency [67].

4.6 Experimental Results

4.6.1 *RQ1. How Accurate Is DDASR in Comparison with the State-of-the-Art Approaches?*

Motivation. A crucial evaluating factor for API sequence recommendation is accuracy. In DDASR, we have incorporated a treatment for tail APIs. We evaluate the accuracy of DDASR concerning baselines on the original Java dataset.

Approach. We first train DDASR on RNN encoder-decoder, Transformer encoder-decoder, and different LLM encoder-decoder architectures as well as DeepAPI with the aforementioned settings on the original Java dataset. For CodeBERT and CodeTrans, we directly utilize their publicly available models to generate API sequences. For BIKER, an information retrieval-based approach, we execute the open source code after necessary modifications. Originally, BIKER calculates the similarity between a new query and existing questions in a historical repository, as well as the similarity between the new query and the official documentation descriptions of APIs in the repository, using a harmonic mean as the ranking score. For API sequence recommendation, we modify the approach to compute the similarity between the new query and the questions in the historical repository, as well as the similarity between the new query and the official documentation descriptions of API sequences in the repository, again using a harmonic mean as the ranking score. Specifically, we calculate the similarity between the new query and each API's official documentation description within an API sequence and use the average of these similarities as the similarity measure between the new query and the official documentation descriptions of API sequences. This tailored approach enhances BIKER's ability to recommend sequences of APIs that are relevant to the new query. For DGAS, we replicate the algorithm and parameters as described in the article. We successfully replicated the results of BIKER and DGAS as reported in their original works [26, 72] to ensure the correctness of our implementation. For GPT-3.5 and GPT-4, we apply nucleus sampling with the sampling parameters set to $top\_p = 0.9$ and $temperature = 0.9$. We use the prompt "*Here are some examples: \n Input Query :* EI. *\n Output API Sequence :* EO. *\n Please generate the top_*K *API sequences in* PL *following the output format in examples that can solve the query :* Q.," where EI represents the example of input query, EO represents the example of output API sequence, K represents the number of API sequences recommended, PL represents the programming language, and Q represents the query of the developer, as the input for GPT-3.5 and GPT-4 to generate API sequences.

We assess the performance using accuracy metrics, specifically BLEU_P, ROUGE_P, MAP_P, and NDCG_P, across the top-1, top-5, and top-10 recommendation results. Additionally, we evaluate the top-1, top-5, and top-10 recommendation results using restored veritable accuracy metrics BLEU_O, ROUGE_O, MAP_O, and NDCG_O.

We perform *t-test*, a statistical test that tests for differences in means between two groups are used to find if differences in reported performance are statistically significant. We first examine the significant differences between DDASR with RNN encoder-decoder, Transformer encoder-decoder, and different LLM encoder-decoder architectures and the best-performing baseline. We utilize *t-test* to test the statistical differences among top-1, top-5, and top-10 results. In addition, the significant differences among DDASR with different LLMs when applying the LLM encoder-decoder architecture are also assessed.

Moreover, we conduct ablation studies by comparing the veritable accuracy of CodeBERT, DASR, DASR+PGT, and DDASR at top-10 recommendation results to evaluate the performance of each component within DDASR. In the experiment, we select CodeBERT as the encoder for DASR, DASR+PGT, and DDASR.

Results. Figure 7 shows an example of the input prompt and output of GPT-4. Tables 4 and 5 show the evaluation results of falsified accuracy and veritable accuracy, respectively, on the original dataset. The bold numbers in the table represent the optimal outcomes. The p-value is the result of the *t-test* between DDASR and the best-performing baseline. Figure 8 shows the comparison of top-k results and the comparison of DDASR using different LLMs. Table 6 shows the evaluation results of ablation studies in veritable accuracy on the original dataset.

Input Prompt

Here are some examples: K

Input Query: Return string representation component. EI PL

Output API Sequence: StringBuffer.<init> StringBuffer.append StringBuffer.toString EO

Please generate the top 5 API sequences in Java following the output format in examples that can solve the query: build json string fields values given module. Q

Output

API Sequence 1: JMenuItem.<init> JMenuItem.setText

API Sequence 2: JMenuItem.<init> JMenu.add

API Sequence 3: JMenuItem.<init> JMenuItem.addActionListener JMenuItem.setActionCommand

API Sequence 4: JMenuItem.<init> JMenuItem.setAccelerator JMenu.add

API Sequence 5: JMenu.<init> JMenuItem.<init> JMenu.add JMenuBar.add

Ground Truth

JMenu.instance JCheckBoxMenuItem.<init> JCheckBoxMenuItem.setName JCheckBoxMenuItem.addActionListener

Fig. 7. An example of the input prompt and output of GPT-4.

Table 4. Evaluation Results in Falsified Accuracy on the Original Java Dataset

	Models	$BLEU_P$			$ROUGE_P$			MAP_P			$NDCG_P$			p
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	
Baselines	DeepAPI	0.2818	0.3659	0.3738	0.4012	0.5260	0.5387	0.3672	0.4308	0.4374	0.3672	0.3572	0.3545	-
	BIKER	0.0000	0.0000	0.0000	0.1201	0.1622	0.1811	0.2892	0.3471	0.4512	0.0700	0.0612	0.0588	-
	DGAS	0.0101	0.0377	0.0635	0.0795	0.1903	0.2424	0.0242	0.0896	0.1493	0.0242	0.0448	0.0209	-
	CodeBERT	0.4024	0.4677	0.4815	0.5436	0.5936	0.6413	0.4774	0.5463	0.5618	0.4774	0.3753	0.3522	-
	CodeTrans	0.4178	0.4708	0.5006	0.5642	0.6266	0.6571	0.5049	0.5750	0.6115	0.5049	0.3655	0.3420	-
	GPT-3.5	0.0692	0.0706	0.0739	0.0899	0.0925	0.0947	0.0687	0.0691	0.0706	0.0763	0.0758	0.0751	-
	GPT-4	0.0698	0.0712	0.0745	0.0914	0.0931	0.0959	0.0699	0.0718	0.0725	0.0770	0.0765	0.0763	-
DDASR	DDASR(RNN)	0.6510	0.7601	0.7825	0.7406	0.8470	0.8653	0.7860	0.8521	0.8684	0.7860	0.7489	0.7316	***
	DDASR(Transformer)	0.6575	0.7685	0.7895	0.7466	0.8501	0.8699	0.7970	0.8632	0.8788	0.7970	0.7903	0.7876	***
	DDASR(CodeBERT)	**0.6682**	0.7707	0.8205	**0.7591**	0.8523	0.8863	**0.8158**	0.8737	0.8971	**0.8158**	**0.8123**	**0.8104**	***
	DDASR(GraphCodeBERT)	0.6628	0.7742	0.8218	0.7563	0.8551	0.8892	0.8135	**0.8784**	**0.8998**	0.8135	0.8109	0.8093	***
	DDASR(PLBART)	0.6649	0.7703	0.8191	0.7582	0.8519	0.8822	0.8129	0.8751	0.8978	0.8129	0.8080	0.8062	***
	DDASR(CodeT5)	0.6484	0.7461	0.8027	0.7221	0.8379	0.8715	0.8026	0.8604	0.8871	0.8026	0.8000	0.8000	***
	DDASR(UniXcoder)	0.6665	**0.7776**	**0.8246**	0.7586	**0.8591**	**0.8906**	0.8137	0.8765	0.8993	0.8137	0.8092	0.8070	***

"***" represents a highly significant statistical difference ($p < 0.001$), "-" represents that there is no value in this cell, and the bold text represents the best performance.

(1) *Compared to the baselines, DDASR demonstrates significant accuracy.* According to Table 4, DDASR significantly improves all metrics ($p < 0.05$), regardless of which architecture it uses. The accuracy of top-1, top-5, and top-10 results is more than 55.19%, 58.47%, and 56.31% on $BLEU_P$, 27.99%, 33.72%, and 32.63% on $ROUGE_P$, 55.67%, 48.19%, and 42.01% on MAP_P, as well as 55.67%, 99.55%, and 106.38% on $NDCG_P$, respectively, which shows DDASR can provide correct API sequences in most cases. The $BLEU_P$ and $BLEU_O$ of BIKER are about 0, indicating that the API sequences recommended by BIKER do not retain the correct call relationship. In addition, CodeBERT performs best as the encoder of DDASR in terms of accuracy in the top-1 recommendation results. In the top-5 and top-10 recommendation results, UniXcoder excels in terms of $BLEU_P$ and $ROUGE_P$ metrics performance, and GraphCodeBERT, on the other hand, demonstrates the best performance in terms of MAP_P metric, and when it comes to

Table 5. Evaluation Results in Veritable Accuracy on the Original Java Dataset

	Models	$BLEU_O$			$ROUGE_O$			MAP_O			$NDCG_O$			p
		top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10	
Baselines	DeepAPI	0.2806	0.3598	0.3701	0.3961	0.5211	0.5332	0.3593	0.4201	0.4269	0.3588	0.3463	0.3461	-
	BIKER	0.0000	0.0000	0.0000	0.1201	0.1622	0.1811	0.2892	0.3471	0.4512	0.0700	0.0612	0.0588	-
	DGAS	0.0101	0.0377	0.0635	0.0795	0.1903	0.2424	0.0242	0.0896	0.1493	0.0242	0.0448	0.0209	-
	CodeBERT	0.3964	0.4631	0.4774	0.5384	0.5886	0.6365	0.4687	0.5371	0.5511	0.4681	0.3671	0.3438	-
	CodeTrans	0.4118	0.4678	0.4967	0.5591	0.6214	0.6527	0.4949	0.5643	0.6023	0.4951	0.3568	0.3333	-
	GPT-3.5	0.0692	0.0706	0.0739	0.0899	0.0925	0.0947	0.0687	0.0691	0.0706	0.0763	0.0758	0.0751	-
	GPT-4	0.0698	0.0712	0.0745	0.0914	0.0931	0.0959	0.0699	0.0718	0.0725	0.0770	0.0765	0.0763	-
DDASR	DDASR(RNN)	0.6476	0.7564	0.7802	0.7354	0.8421	0.8603	0.7771	0.8438	0.8591	0.7763	0.7396	0.7224	***
	DDASR(Transformer)	0.6551	0.7613	0.7854	0.7412	0.8451	0.8647	0.7881	0.8530	0.8689	0.7888	0.7800	0.7781	***
	DDASR(CodeBERT)	**0.6612**	0.7654	0.8185	**0.7541**	0.8479	0.8817	**0.8059**	0.8647	0.8891	**0.8061**	**0.8036**	**0.8007**	***
	DDASR(GraphCodeBERT)	0.6564	0.7685	0.8197	0.7521	0.8500	0.8843	0.8047	**0.8691**	**0.8900**	0.8043	0.8004	0.8001	***
	DDASR(PLBART)	0.6587	0.7653	0.8154	0.7529	0.8467	0.8774	0.8040	0.8662	0.8898	0.8022	0.7972	0.7983	***
	DDASR(CodeT5)	0.6447	0.7435	0.7966	0.7167	0.8322	0.8661	0.7933	0.8498	0.8781	0.7936	0.7889	0.7908	***
	DDASR(UniXcoder)	0.6598	**0.7741**	**0.8215**	0.7533	**0.8546**	**0.8857**	0.8048	0.8677	0.8900	0.8027	0.7991	0.7989	***

"***" represents a highly significant statistical difference ($p < 0.001$), "-" represents that there is no value in this cell, and the bold text represents the best performance.

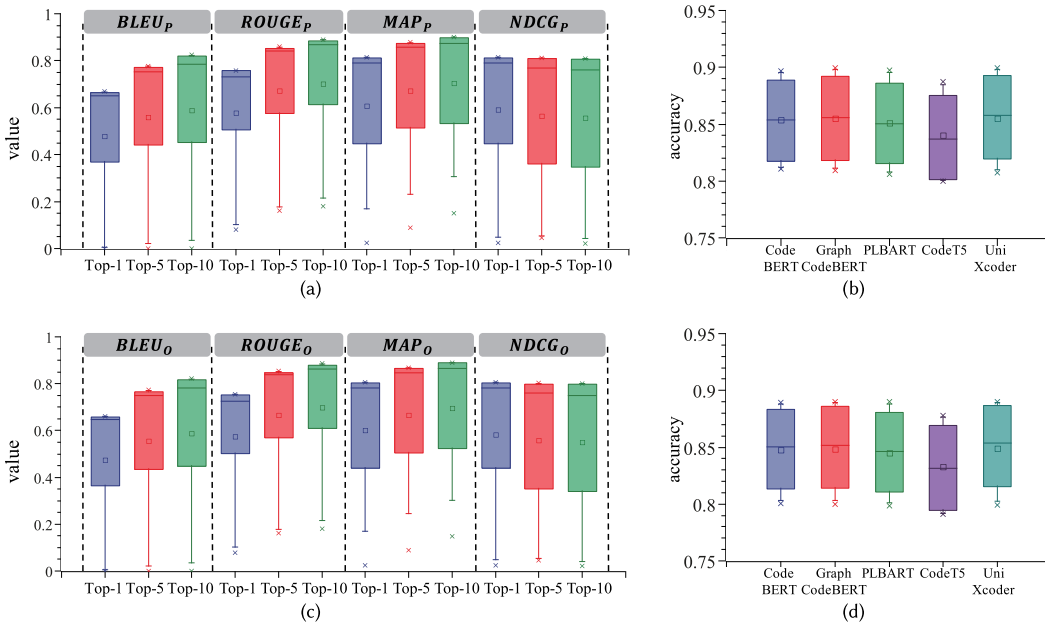

Fig. 8. Accuracy comparative analysis on the original Java dataset. (a) Comparison of top-k results under different falsified accuracy metrics, (b) comparison of DDASR with different LLMs in falsified accuracy, (c) comparison of top-k results under different veritable accuracy metrics, and (d) comparison of DDASR with different LLMs in veritable accuracy.

$NDCG_P$ metric, CodeBERT outperforms other encoders. From Table 5, we find that there are consistent results in terms of veritable accuracy. From Figure 8(b) and (d), it can be observed that DDASR using GraphCodeBERT demonstrates the best overall performance. Moreover, it is observed that GPT-3.5 and GPT-4 exhibit sub-optimal performance. We hypothesize that, due to the high degree of flexibility in LLMs when generating output, while some of the API sequences generated may be semantically correct, they may differ from the ground truth

Table 6. Evaluation Results of Ablation Studies in Veritable Accuracy on the Original Dataset

Models	$BLEU_O$	$ROUGE_O$	MAP_O	$NDCG_O$
CodeBERT	0.4774	0.6365	0.5511	0.3438
DASR(CodeBERT)	**0.8375**	**0.9098**	**0.8972**	**0.8184**
DASR(CodeBERT)+PGT	0.7901	0.8522	0.8654	0.7963
DDASR(CodeBERT)	0.8135	0.8817	0.8891	0.8007

The bold text represents the best performance.

as shown in Figure 7. This discrepancy results in poor performance when evaluated using metrics.

(2) *An increase in the number of recommended results tends to result in improved accuracy*. As the number of recommendation results rises, $BLEU_P$, $BLEU_O$, $ROUGE_P$, $ROUGE_O$, MAP_P, and MAP_O increase. It indicates that developers are more likely to discover the optimal solution when more results are offered. Nevertheless, the growth is not significant ($p > 0.05$). In addition, $NDCG_P$ and $NDCG_O$ decrease as the number of recommendation results increases. It verifies that the first appearance of correct APIs is concentrated in the position in front of the recommendation results.

(3) *The performance gap of DDASR with the LLM encoder-decoder architecture is not very significant when applying different LLMs ($p > 0.05$).* In the top-10 recommended results, the differences of DDASR with different LLMs in $BLEU_P$, $ROUGE_P$, MAP_P, and $NDCG_P$ are only 0.67%, 2.19%, 1.43%, and 1.30%, respectively, and in $BLEU_O$, $ROUGE_O$, MAP_O, and $NDCG_O$ are only 3.13%, 2.26%, 1.36%, and 1.25%, respectively.

(4) *Merging API fragments into complete APIs can effectively increase the accuracy*. From Table 6, we observe that DASR, which only merges API fragments into complete APIs, performs the best. Adding the pseudo ground truth construction component to DASR causes a significant drop in accuracy. Our method, DDASR, achieves a balance between accuracy and diversity. Although its accuracy slightly decreases compared to DASR, it still shows a significant improvement compared to the baselines.

Answer to RQ1: When using alternative encoder-decoder architectures on the original Java dataset, DDASR consistently outperforms the baseline models in terms of accuracy. Furthermore, the performance of DDASR with different LLM encoder-decoder models is not pronounced.

4.6.2 RQ2. Does DDASR Improve Diversity While Maintaining Accuracy?

Motivation. Diversity is another essential factor in API sequence recommendation, which is disregarded by previous approaches. In RQ2, we look at how well our DDASR performs in terms of diversity. Furthermore, accuracy and diversity are frequently incompatible objectives. We also investigate how DDASR affects accuracy when improving diversity.

Approach. To answer this question, we conduct comparative experiments at the top-10 recommendation results on the diverse Java dataset and the Python dataset. We train DDASR and DASR on RNN encoder-decoder, Transformer encoder-decoder, and different LLM encoder-decoder architectures as well as deep learning-based baselines with the aforementioned settings on the diverse Java dataset and the Python dataset, respectively. Specifically, when working on the Python dataset, we directly utilize the publicly available model of CodeBERT to generate results. The execution of BIKER, DGAS, GPT-3.5, and GPT-4 is the same as RQ1.

Table 7. Evaluation Results in Accuracy and Diversity at Top-10 Recommendation Results on the Diverse Java Dataset

	Models	$BLEU_P$	$BLEU_O$	$ROUGE_P$	$ROUGE_O$	MAP_P	MAP_O	$NDCG_P$	$NDCG_O$	Coverage	p
Baselines	DeepAPI	0.0943	0.0941	0.1567	0.1532	0.1815	0.1804	0.1506	0.1401	0.0329	-
	BIKER	0.0000	0.0000	0.0235	0.0235	0.0554	0.0554	0.0039	0.0039	0.0319	-
	DGAS	0.0081	0.0081	0.2039	0.2039	0.0176	0.0176	0.0019	0.0019	0.0815	-
	CodeBERT	0.0464	0.0402	0.1222	0.1198	0.0909	0.0896	0.0347	0.0309	0.0544	-
	CodeTrans	0.2477	0.2389	0.3743	0.3695	0.3375	0.3246	0.1324	0.1276	0.0744	-
	GPT-3.5	0.0462	0.0462	0.0498	0.0498	0.0543	0.0543	0.0527	0.0527	0.0036	-
	GPT-4	0.0449	0.0449	0.0501	0.0501	0.0549	0.0549	0.0534	0.0534	0.0035	-
DASR	DASR(RNN)	0.2926	0.2865	0.4152	0.4038	0.4557	0.4442	0.4243	0.4168	0.0400	***
	DASR(Transformer)	0.4124	0.4027	0.5690	0.5603	0.4865	0.4796	0.4268	0.4127	0.0406	***
	DASR(CodeBERT)	**0.5260**	**0.5086**	**0.6387**	**0.6174**	0.6231	0.6136	**0.4752**	**0.4621**	0.0430	***
	DASR(GraphCodeBERT)	0.4711	0.4609	0.5782	0.5680	0.5885	0.5801	0.4119	0.4093	0.0418	***
	DASR(PLBART)	0.4769	0.4688	0.5801	0.5769	0.5876	0.5796	0.4249	0.4117	0.0438	***
	DASR(CodeT5)	0.4992	0.4911	0.6038	0.5967	0.6007	0.5943	0.4521	0.4474	0.0446	***
	DASR(UniXcoder)	0.4851	0.4824	0.5866	0.5851	0.6056	0.5994	0.4208	0.4165	0.0428	***
DDASR	DDASR(RNN)	0.3221	0.3196	0.4471	0.4400	0.4493	0.4438	0.4526	0.4435	0.1064	***
	DDASR(Transformer)	0.4002	0.3971	0.5561	0.5513	0.4968	0.4883	0.4344	0.4263	0.0758	***
	DDASR(CodeBERT)	0.5159	0.5077	0.6262	0.6149	**0.6365**	**0.6255**	0.4629	0.4554	**0.1085**	***
	DDASR(GraphCodeBERT)	0.4629	0.4581	0.5689	0.5675	0.5845	0.5769	0.4115	0.4073	0.1038	***
	DDASR(PLBART)	0.4550	0.4516	0.5671	0.5664	0.5832	0.5770	0.4067	0.4037	0.0996	***
	DDASR(CodeT5)	0.4880	0.4826	0.5914	0.5852	0.6100	0.6012	0.4404	0.4343	0.0978	***
	DDASR(UniXcoder)	0.4759	0.4702	0.5796	0.5780	0.6018	0.5927	0.4199	0.4154	0.1003	***

"***" represents a highly significant statistical difference ($p < 0.001$), "-" represents that there is no value in this cell, and the bold text represents the best performance.

We calculate the accuracy metrics of $BLEU_P$, $ROUGE_P$, MAP_P, and $NDCG_P$, the veritable accuracy metrics of $BLEU_O$, $ROUGE_O$, MAP_O, and $NDCG_O$, as well as the diversity metric of coverage on the top-10 results of all models. Moreover, we conduct significance analysis for performance differences between DASR and the best-performing baseline, between DDASR and the best-performing baseline, among DASR with different LLM encoder-decoder architectures, among DDASR with different LLM encoder-decoder architectures, between DASR and DDASR on coverage, and between DASR and DDASR on accuracy.

We conduct ablation studies by comparing the veritable accuracy of CodeBERT, DASR, DASR+PGT, and DDASR at top-10 recommendation results to evaluate the performance of each component within DDASR on the diverse Java dataset and the Python dataset. In the experiment, we select CodeBERT as the encoder for DASR, DASR+PGT, and DDASR. Additionally, we divide both the test sets of the diverse Java dataset and the Python dataset into tail API datasets, a subset of queries whose ground truth contains tail APIs, and non-tail API datasets, a subset of queries whose ground truth contains non-tail APIs only. We conduct ablation studies separately on the non-tail and tail API datasets within both the diverse Java dataset and the Python dataset.

Results. Table 7 shows the experimental results on the diverse Java dataset. Figure 9 shows the change of accuracy metrics in DASR and DDASR with epoch during the training phase. Among them, DASR and DDASR employ seven encoder-decoder models. The performance results on the Python dataset are shown in Table 8. The p-value in Tables 7 and 8 is the result of the *t-test* between DASR and the best-performing baseline, as well as between DDASR and the best-performing baseline. Figure 10 shows the comparison of DASR using different LLMs and the comparison of DDASR using different LLMs on the diverse Java dataset. Figure 11 shows the comparison of DASR and DDASR on accuracy and the comparison of DASR and DDASR on coverage. Figure 12 shows

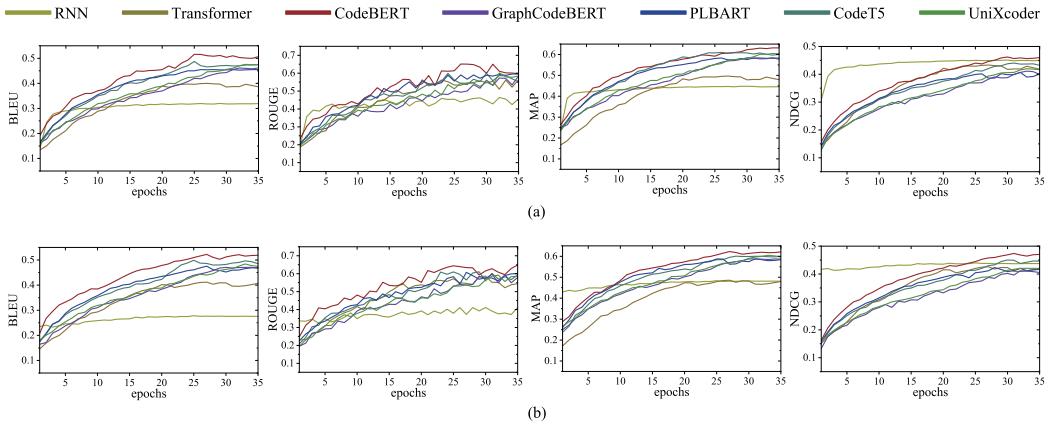

Fig. 9. The comparison of BLEU, ROUGE, MAP, and NDCG for each epoch on the diverse Java dataset. (a) The performance of DDASR, and (b) the performance of DASR.

Table 8. The Performance in Accuracy and Diversity at Top-10 Recommendation Results on the Python Dataset

	Models	$BLEU_P$	$BLEU_O$	$ROUGE_P$	$ROUGE_O$	MAP_P	MAP_O	$NDCG_P$	$NDCG_O$	Coverage	p
Baselines	DeepAPI	0.4386	0.4227	0.6561	0.6403	0.5012	0.4835	0.3561	0.3342	0.1851	-
	BIKER	0.1865	0.1865	0.2075	0.2075	0.4037	0.4037	0.3158	0.3158	0.1164	-
	DGAS	0.4423	0.4423	0.3466	0.3466	0.5241	0.5241	0.3455	0.3455	0.1508	-
	CodeBERT	0.4099	0.3896	0.6113	0.6002	0.5249	0.5041	0.1836	0.1702	0.2080	-
	CodeTrans	0.3998	0.3854	0.6045	0.5987	0.5176	0.4988	0.1901	0.1819	0.1924	-
	GPT-3.5	0.0753	0.0753	0.0899	0.0899	0.1396	0.1396	0.1303	0.1303	0.0248	-
	GPT-4	0.0872	0.0872	0.0981	0.0981	0.1507	0.1507	0.1443	0.1443	0.0259	-
DASR	DASR(RNN)	0.4843	0.4754	0.6712	0.6693	0.5440	0.5229	0.3767	0.3584	0.1475	ns
	DASR(Transformer)	0.4982	0.4799	0.6828	0.6713	0.5845	0.5770	0.2752	0.2698	0.2320	ns
	DASR(CodeBERT)	**0.5432**	**0.5240**	**0.7014**	**0.6933**	**0.6117**	**0.6032**	0.3660	0.3587	0.1954	ns
	DASR(GraphCodeBERT)	0.5058	0.4896	0.6837	0.6830	0.5781	0.5704	0.3206	0.3146	0.1786	ns
	DASR(PLBART)	0.5098	0.4927	0.6858	0.6781	0.5860	0.5810	0.3234	0.3150	0.1784	ns
	DASR(CodeT5)	0.5389	0.5228	0.6983	0.6926	0.6090	0.6017	0.3633	0.3572	0.1736	ns
	DASR(UniXcoder)	0.4885	0.4719	0.6746	0.6689	0.5682	0.5610	0.2881	0.2824	0.1946	ns
DDASR	DDASR(RNN)	0.4560	0.4507	0.6658	0.6632	0.5219	0.5134	0.3211	0.3117	0.2232	ns
	DDASR(Transformer)	0.4292	0.4128	0.6453	0.6227	0.5108	0.5039	0.2546	0.2405	**0.2378**	ns
	DDASR(CodeBERT)	0.4858	0.4679	0.6722	0.6678	0.5672	0.5592	0.3677	0.3605	0.2247	ns
	DDASR(GraphCodeBERT)	0.4928	0.4768	0.6781	0.6710	0.5726	0.5620	**0.3879**	**0.3809**	0.2103	ns
	DDASR(PLBART)	0.4869	0.4738	0.6733	0.6682	0.5682	0.5619	0.3609	0.3558	0.2216	ns
	DDASR(CodeT5)	0.4544	0.4453	0.6650	0.6611	0.5295	0.5290	0.3603	0.3543	0.2256	ns
	DDASR(UniXcoder)	0.4984	0.4828	0.6829	0.6719	0.5776	0.5668	0.3843	0.3758	0.2177	ns

"ns" represents no significance ($p > 0.05$), "-" represents that there is no value in this cell, and the bold text represents the best performance.

the comparison of DASR and DDASR with different LLMs on accuracy on the diverse Java dataset. Figures 13–15 show the comparison on the Python dataset. Table 9 shows the evaluation results of ablation studies on the complete, non-tail, and tail diverse Java dataset and on the complete, non-tail, tail Python dataset.

(1) *Both DASR and DDASR exhibit significant accuracy gains over baselines ($p < 0.05$), and the difference between DASR and DDASR is not particularly noticeable ($p > 0.05$) on the diverse Java dataset*. As can be seen from Table 7, accuracy metrics for both DASR and DDASR

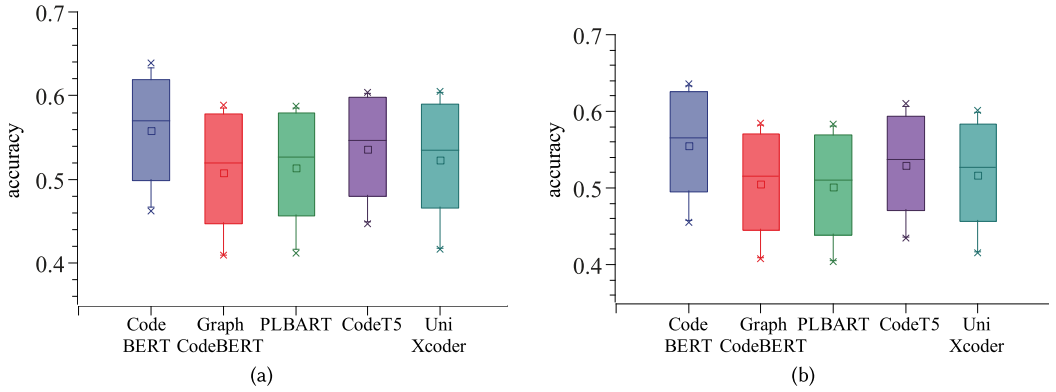

Fig. 10. Accuracy comparative analysis of DASR and DDASR applying different LLMs on the diverse Java dataset. (a) Comparison of DASR with different LLMs applied, and (b) comparison of DDASR with different LLMs.

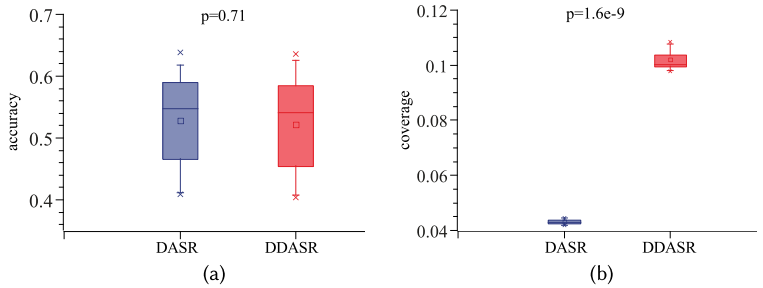

Fig. 11. Comparative analysis between DASR and DDASR on the diverse Java dataset. (a) Comparison on accuracy, and (b) comparison on coverage.

have significantly increased in comparison to the baselines ($p < 0.05$), which illustrates that word embedding of the complete API outperforms character embedding of API fragments. Regardless of which LLMs DDASR uses, DDASR exceeds baselines with the best performance by more than 83.69%, 89.03%, 51.51%, 53.29%, 77.73%, 101.54%, 109.77%, and 188.15% on all accuracy metrics. From Figure 9, we can observe that DASR and DDASR using CodeBERT as the encoder consistently outperform other models. As shown in Figure 11(a), the difference is not particularly apparent ($p > 0.05$) between DASR and DDASR. Moreover, for DDASR and DASR, which use the same LLMs, DDASR has not shown a significant decrease in accuracy, and in some cases, it even outperforms DASR as shown in Figure 12. It indicates that the treatment of tail APIs in DDASR does not cause a significant drop in the accuracy of recommendation results.

(2) *Both DASR and DDASR with LLMs perform better than baselines in terms of accuracy, and the differences are not particularly noticeable ($p > 0.05$) on the Python dataset*. Regardless of which LLMs DDASR uses, DDASR exceeds baselines with the best performance by more than 2.74%, 0.68%, 1.36%, 3.25%, 4.94%, 1.20%, 2.47%, and 6.01% on all accuracy metrics, although the difference is not significant ($p > 0.05$). We speculate it is due to the more straightforward composition of Python APIs. For example, "*scipy.special.digamma*" is a typical API in the Python dataset. There are essentially no symbols like "<" and ">" in API, and no nested usage patterns as in the Java dataset, making the effect of merging APIs less obvious than

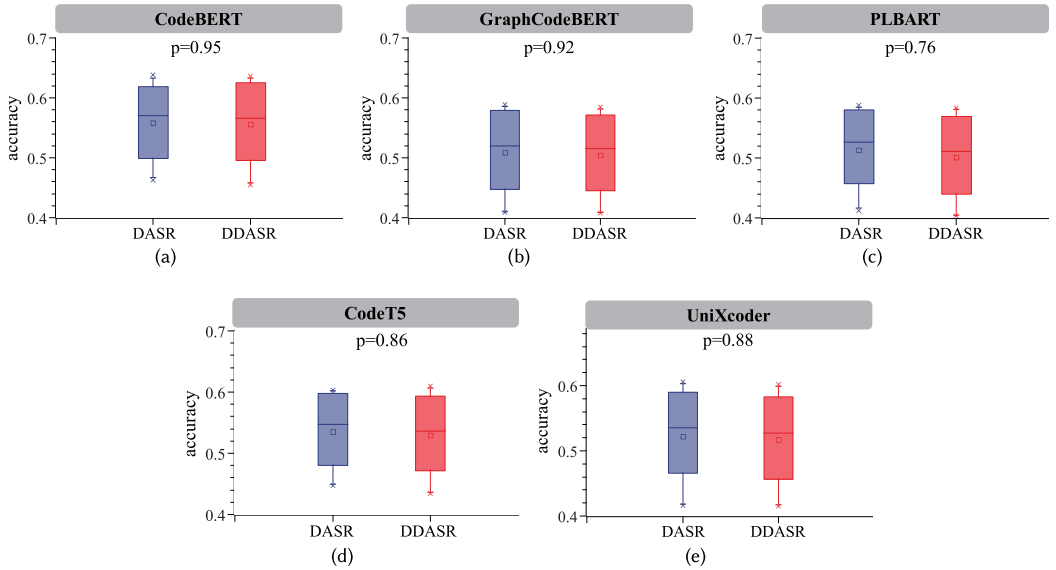

Fig. 12. Accuracy comparative analysis between DASR and DDASR with different LLMs on the diverse Java dataset. (a) Comparison with CodeBERT, (b) comparison with GraphCodeBERT, (c) comparison with PLBART, (d) comparison with CodeT5, and (e) comparison with UniXcoder.

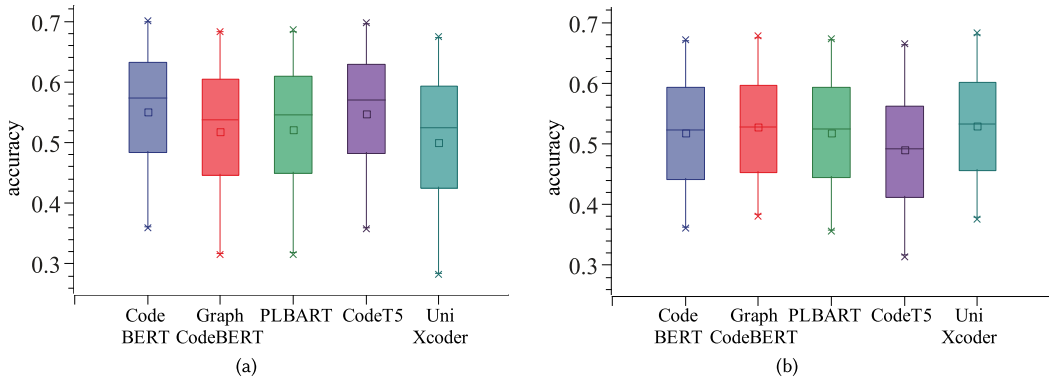

Fig. 13. Accuracy comparative analysis of DASR and DDASR applying different LLMs on the Python dataset. (a) Comparison of DASR with different LLMs, and (b) comparison of DDASR with different LLMs.

in the Java dataset. Furthermore, it can be observed from Figures 14(a) and 15 that, on the Python dataset, the accuracy difference between DASR and DDASR is not significant.

(3) *On both the diverse Java dataset and the Python dataset, the restored veritable accuracy is lower than the falsified accuracy, but DASR and DDASR still perform better than baselines.* Compared to BLEU_P, ROUGE_P, MAP_P, and NDCG_P, BLEU_O, ROUGE_O, MAP_O, and NDCG_O all show a decrease. The reduction for DASR and deep learning-based baselines, however, is frequently more than for DDASR. Compared to the falsified accuracy, DDASR, DASR, and deep learning-based baselines have, on average, decreased the restored veritable accuracy by 1.17%, 1.75%, and 4.17% on the diverse Java dataset, as well as decreased by 1.97%, 2.01%,

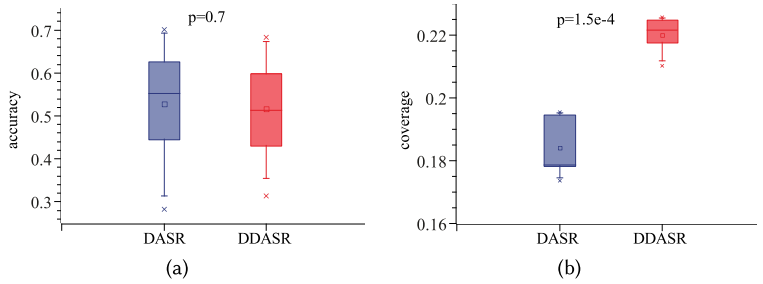

Fig. 14. Comparative analysis between DASR and DDASR on the Python dataset. (a) Comparison on accuracy, and (b) comparison on coverage.

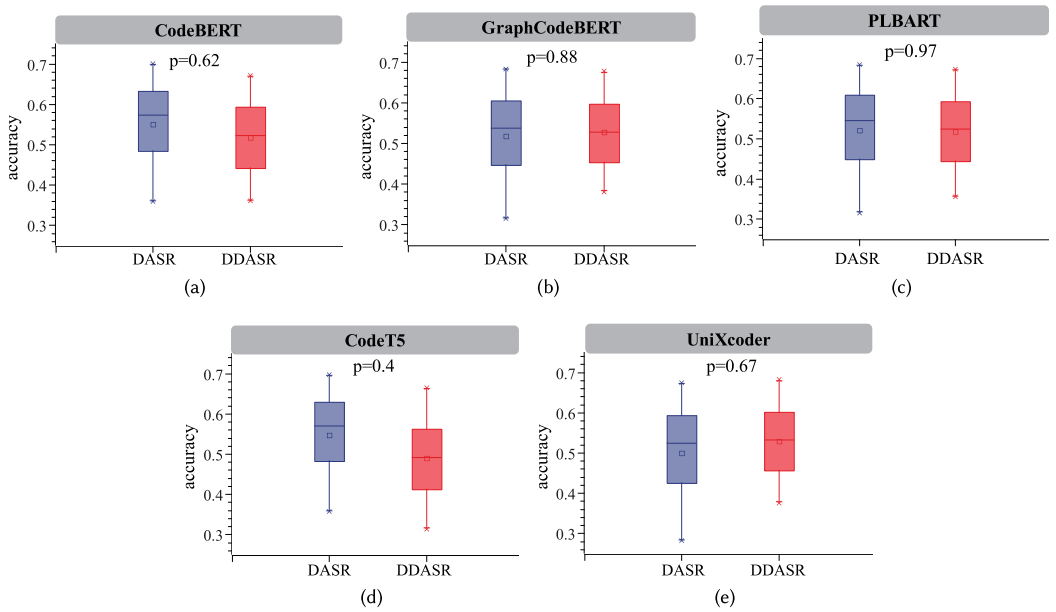

Fig. 15. Accuracy comparative analysis between DASR and DDASR with different LLMs on the Python dataset. (a) Comparison with CodeBERT, (b) comparison with GraphCodeBERT, (c) comparison with PLBART, (d) comparison with CodeT5, and (e) comparison with UniXcoder.

and 3.85% on the Python dataset, respectively. This suggests that constructing the pseudo ground truth through clustering and relocating is more effective than using *<UNK>* tags for replacement.

(4) *The differences among DDASR with different LLMs and among DASR applying different LLMs are not significant on accuracy ($p > 0.05$).* We evaluate the statistical differences among DASR with different LLM encoder-decoder architectures and among DDASR with different LLM encoder-decoder architectures. It can be found that there are no significant differences ($p > 0.05$) as shown in Figures 10 and 13. Furthermore, the results suggest that DASR and DDASR, when utilizing CodeBERT, perform the best on the diverse Java dataset, while DASR using CodeBERT and DDASR using UniXcoder excel on the Python dataset.

(5) *DDASR significantly enhances diversity ($p < 0.05$) while maintaining accuracy ($p > 0.05$).* As shown in Tables 7 and 8, DDASR demonstrates a significant improvement in both diversity

Table 9. Evaluation Results of Ablation Studies on the Diverse Java Dataset and the Python Dataset

Datasets		Models	$BLEU_O$	$ROUGE_O$	MAP_O	$NDCG_O$	Coverage
Diverse Java dataset	Complete dataset	CodeBERT	0.0402	0.1198	0.0896	0.0309	0.0544
		DASR(CodeBERT)	**0.5086**	**0.6174**	0.6136	**0.4621**	0.0430
		DASR(CodeBERT)+PGT	0.4921	0.6001	0.6142	0.4512	**0.1098**
		DDASR(CodeBERT)	0.5077	0.6149	**0.6255**	0.4554	0.1085
	Non-tail API dataset	CodeBERT	0.0416	0.1244	0.0946	0.0318	-
		DASR(CodeBERT)	**0.5154**	**0.6235**	0.6188	**0.4640**	-
		DASR(CodeBERT)+PGT	0.4909	0.5982	0.6155	0.4515	-
		DDASR(CodeBERT)	0.5105	0.6189	**0.6292**	0.4560	-
	Tail API dataset	CodeBERT	0.0382	0.1101	0.0798	0.0287	-
		DASR(CodeBERT)	0.4958	0.6052	0.6033	**0.4571**	-
		DASR(CodeBERT)+PGT	0.4943	0.6044	0.6110	0.4508	-
		DDASR(CodeBERT)	**0.5010**	**0.6098**	**0.6178**	0.4541	-
Python dataset	Complete dataset	CodeBERT	0.3896	0.6002	0.5041	0.1702	0.2080
		DASR(CodeBERT)	**0.5240**	**0.6933**	**0.6032**	0.3587	0.1954
		DASR(CodeBERT)+PGT	0.4518	0.6542	0.5553	0.3572	**0.2401**
		DDASR(CodeBERT)	0.4679	0.6678	0.5592	**0.3605**	0.2247
	Non-tail API dataset	CodeBERT	0.3902	0.6037	0.5068	0.1734	-
		DASR(CodeBERT)	**0.5289**	**0.6962**	**0.6091**	0.3588	-
		DASR(CodeBERT)+PGT	0.4510	0.6528	0.5559	0.3574	-
		DDASR(CodeBERT)	0.4671	0.6669	0.5598	**0.3606**	-
	Tail API dataset	CodeBERT	0.3772	0.5891	0.4905	0.1638	-
		DASR(CodeBERT)	0.4699	0.6673	0.5578	0.3579	-
		DASR(CodeBERT)+PGT	0.4533	0.6565	0.5548	0.3568	-
		DDASR(CodeBERT)	**0.4703**	**0.6723**	**0.5585**	**0.3600**	-

"-" represents that there is no value in this cell and the bold text represents the best performance.

and accuracy compared to the baselines. Results in Table 9 indicate that DASR, DASR+PGT, and DDASR all show improvements in accuracy over the baselines, with DASR performing the best in terms of accuracy, though its diversity remains low. DASR+PGT offers a substantial increase in diversity but at a significant loss of accuracy compared to DASR. In contrast, DDASR significantly enhances diversity, as illustrated in Figures 11(b) and 14(b), without significantly compromising accuracy, as shown in Figures 12 and 15.

(6) *DDASR excels in recommending tail APIs more accurately*. As shown in Tables 7 and 8, DDASR performs well in terms of coverage, indicating that it can recommend a greater number of tail APIs. In Table 9, DDASR exhibits the best performance on the tail API dataset, demonstrating its ability to recommend more precise tail APIs.

Answer to RQ2: DDASR exhibits superior diversity compared to DASR and the baselines. While DASR+PGT exhibits excellent diversity, it causes a sharp decline in accuracy. In terms of accuracy, there is no significant difference between DDASR and DASR. However, DDASR does outperform the baselines. These findings indicate that our DDASR can significantly increase diversity without compromising accuracy.

4.6.3 *RQ3. Can DDASR Help Programmers Address Tasks More Effectively?*

Motivation. The ultimate goal of API sequence recommendation is to assist developers. Thus, in this section, we conduct a human evaluation to assess our DDASR.

Table 10. The Rating Scores of the Human Evaluation

Models	Evaluation Sets	# Rating Score of Developer 1				# Rating Score of Developer 2				# Rating Score of Developer 3				# Rating Score of Developer 4				# Rating Score of Developer 5				# Rating Score of Developer 6			
		Zero	One	Two	Avg	Zero	One	Two	Avg	Zero	One	Two	Avg	Zero	One	Two	Avg	Zero	One	Two	Avg	Zero	One	Two	Avg
DeepAPI	Tail API set	13	5	0	0.28	13	5	0	0.28	15	3	0	0.17	10	7	1	0.5	11	7	0	0.39	13	5	0	0.28
	Non-tail API set	36	40	6	0.63	34	44	4	0.63	39	38	5	0.59	34	43	5	0.65	38	40	4	0.59	46	33	3	0.48
BIKER	Tail API set	13	5	0	0.28	12	6	0	0.33	13	5	0	0.28	10	8	0	0.44	12	5	1	0.28	10	6	2	0.33
	Non-tail API set	46	30	6	0.52	40	38	4	0.56	53	27	2	0.38	40	35	7	0.59	40	36	6	0.59	43	30	9	0.59
DGAS	Tail API set	10	8	0	0.44	11	6	1	0.44	10	7	1	0.5	11	7	0	0.39	12	5	1	0.39	10	6	2	0.56
	Non-tail API set	29	41	12	0.79	34	48	0	0.59	30	39	13	0.79	27	43	12	0.82	29	38	15	0.83	23	32	27	1.05
CodeBERT	Tail API set	6	11	1	0.72	6	10	2	0.78	8	10	0	0.56	8	9	1	0.61	5	9	4	0.94	2	13	3	1.06
	Non-tail API set	25	50	7	0.78	20	50	12	0.90	17	42	23	1.07	23	35	24	0.98	19	55	18	1.11	14	54	14	1.00
CodeTrans	Tail API set	5	13	0	0.72	6	10	2	0.78	3	11	4	1.06	1	12	5	1.22	0	10	8	1.44	1	10	7	1.33
	Non-tail API set	23	55	4	0.77	15	43	24	1.11	16	46	20	1.05	11	54	17	1.07	16	51	15	0.99	12	57	13	1.01
GPT-3.5	Tail API set	11	4	3	0.56	16	2	0	0.11	11	5	2	0.50	14	1	3	0.39	7	7	4	0.83	9	6	3	0.67
	Non-tail API set	36	18	28	0.90	29	18	35	1.07	30	20	32	1.02	24	15	43	1.23	21	21	40	1.23	20	27	35	1.18
GPT-4	Tail API set	10	5	3	0.61	11	3	4	0.61	8	7	3	0.72	13	2	3	0.44	6	8	4	0.89	7	8	3	0.78
	Non-tail API set	30	24	28	0.98	28	17	37	1.11	27	25	30	1.04	23	16	43	1.24	20	20	42	1.27	18	30	34	1.20
DDASR	Tail API set	3	9	6	1.17	4	9	5	1.06	3	8	7	1.22	0	9	9	1.5	0	6	12	1.67	0	8	10	1.56
	Non-tail API set	12	27	43	1.38	18	27	37	1.23	16	20	46	1.37	5	47	30	1.30	2	26	54	1.63	4	24	54	1.61

A score of zero indicates that the generated API sequences are of no help in solving the requirement, a score of one indicates it is valuable to the developer, and a score of two suggests it can solve the requirement. Avg represents the average score given by each developer to each evaluation set.

Approach. We randomly select 100 queries and generate the top-10 API sequences using all baselines and DDASR, respectively, to evaluate the validity and usefulness of the results for programmers. Six developers who study or work in computer-related fields with experience in Java projects are invited to join a subject group. Three of the developers are currently pursuing master's degrees with more than 2 years of experience programming in Java, while the other three are engaged in Java development in Internet companies, such as Alibaba, Byte Dance, and Huawei. It is important to note that they are not co-authors. They independently give a rating score for the generated API sequences according to their experience. A score of zero indicates that they believed the API sequences generated are completely useless for the query, while a score of one denotes they are valuable to the developers. For example, the key API for solving the query appears in the generated API sequences, or the results inspire them. The rating score is two if the API sequences generated are a way to the query. In addition, the developers are also allowed to search the Internet for unfamiliar concepts. Specifically, to ensure fairness in the human evaluation, each developer was provided only the query and the corresponding API sequences generated by different models, without being informed which model produced them. We further subdivide the 100 queries into a tail API set, a subset of queries whose ground truth contains tail APIs, and a non-tail API set, a subset of queries whose ground truth contains non-tail APIs only. We analyze the rating scores given by developers to the results generated by all methods in both sets.

Results. The results of the human evaluation are shown in Table 10. Developers 1–3 are students, and developers 4–6 have working experience in Internet companies. Table 11 shows 10 queries with the same rating scores and their ground truth, top-3 generated API sequences, and human rating scores. These queries vary in length, containing both very long and short queries that were not included in the training set. They include requirements for page layout, mathematical calculation, string processing, file streaming, and so on. In addition, tail APIs are highlighted in italics in the table.

Table 11. Results of Human Evaluation with API Sequences in the Ground Truth, Top-3 Recommendation Results by DDASR, and Human Rating Scores with 10 Queries

ID	Query	API Sequence in the Ground Truth	API Sequences Generated by DDASR	Rating Score
1	Create appropriate menu item	JMenu.instance JCheckBoxMenuItem.<init> JCheckBoxMenuItem.setName JCheckBoxMenuItem.addActionListener	1 JMenuItem.<init> JMenuItem.addActionListener 2 JMenuItem.<init> JMenuItem.setMnemonic JMenuItem.addActionListener 3 JMenuItem.<init> ActionListener.<init> JMenuItem.addActionListener	1
2	Build json string fields values given module	StringBuilder.<init> StringBuilder.append Map<String,List<Field».keySet String.cast	1 String.instance String.cast Collection.instance Collection.cast 2 StringBuilder.<init> StringBuilder.append StringBuilder.deleteCharAt StringBuilder.toString 3 String.instance String.cast Object.toString Collection.instance	0
3	Calculate standard deviation	Math.sqrt Math.round	1 Math.pow Math.sqrt 2 Math.pow Math.round Math.sqrt 3 Math.pow Math.sqrt	2
4	Create full classpath including existing classpath additional paths jars service files	System.getenv StringBuilder.<init> File.getAbsolutePath	1 System.getenv StringBuilder.<init> StringBuilder.append 2 StringBuilder.<init> StringBuilder.append StringBuilder.toString 3 StringBuilder.<init> StringBuilder.append File.getAbsolutePath	2
5	Create returns xml *<UNK>* session element, stores *<UNK>* session details, including *<UNK>* state	Document.createElement Element.setAttribute Integer.toString	1 Element.<init> Element.setAttribute Element.addContent 2 Element.<init> String.valueOf Element.setAttribute 3 Element.<init> Element.setAttribute Element.<init>	1
6	End operation	Map<Integer,LinkedList<Operation».get LinkedList<Operation>.isEmpty LinkedList<Operation>.getLast	1 Map<Integer,LinkedList<Operation».get LinkedList<Operation>.getLast LinkedList<Operation>.removeLast 2 Map<Integer,LinkedList<Operation».get LinkedList<Operation>.isEmpty LinkedList<Operation>.getLast 3 Map<Integer,LinkedList<Operation».get LinkedList<Operation>.isEmpty LinkedList<Operation>.isEmpty	2
7	Open connection given url specific http method	URL.openConnection HttpURLConnection.cast HttpURLConnection.setDoOutput	1 URL.<init> URL.openConnection HttpURLConnection.cast 2 URL.openConnection HttpURLConnection.cast HttpURLConnection.setRequestProperty 3 URL.<init> URL.getHost URL.<init>	2
8	Return elements contained element given fully qualified name	*DTMNodeProxy.getElementsByTagNameNS* QName.getName ArrayList.<init> NodeList.item	1 *DTMNodeProxy.getElementsByTagName* ArrayList.<init> NodeList.getLength NodeList.item 2 *DTMNodeProxy.getElementsByTagName* ArrayList.<init> Element.instance Element.cast 3 *DTMNodeProxy.getElementsByTagName* ArrayList.<init> NodeList.item List.add	1
9	Locate visible view latest adapter position	LayoutManager.getChildCount *LayoutManager.getChildAt* View.getLayoutParams	1 *LayoutManager.getChildAt* Location.getX 2 *LayoutManager.getChildAt* *LayoutManager.getChildAt* 3 *LayoutManager.getChildAt* Location.getX Location.getY	1
10	Set right child indent	*AbstractLayoutCache.invalidateSizes* BasicTreeUI.updateSize	1 BasicTreeUI.updateSize ArrayList.add 2 BasicTreeUI.updateSize 3 *AbstractLayoutCache.invalidateSizes*	2

(1) *DDASR is effective in recommending appropriate API sequences for developers in the majority of cases*. Results in Table 10 show that the average scores given by the six developers for results generated by DDASR are higher than results generated by baselines. Moreover, the six developers believe that the results generated by DDASR could either directly solve the query problems or provide substantial help (rating scores greater than or equal to one) in a higher proportion of all evaluated queries (over 78%). This indicates that DDASR can effectively recommend suitable API sequences to developers. Notably, compared to developers 1–3, developers 4–6 give fewer rating scores of zero and higher average scores. This difference is likely due to the higher programming expertise of the first group, allowing them to more effectively identify useful APIs. Specifically, for instance, GPT-3.5 and GPT-4 may generate seemingly reasonable but incorrect APIs by concatenating keywords from the query into API names. This can easily mislead developers with less programming experience, while those with more experience are better at identifying these erroneous APIs.

(2) *DDASR can generate more accurate API sequences for length queries, while its performance may degrade if the query contains infrequent words represented as <UNK> tags*. DDASR recommends correct results for length queries like *Query 4*. *Query 5* is also lengthy, with only one API hitting, which cannot solve the query exhaustively. We assume this is because some words, which occur less frequently in the query, are replaced with *<UNK>* tags during data pre-processing. These words may be domain-specific keywords, thus preventing mining the true intent of the query when training.

(3) *For short queries, DDASR tends to generate correct results, especially when the query contains words related to the API sequence or the topic of the query is precise*. The generated API sequences will get high accuracy if the query contains words related to the API sequence, like the class name. *Query 6*, "*end operation*" contains the "*Operation*" class in the API sequence of the ground truth, and the generated API sequences contain all APIs in the first two of results. For queries that focus on a specific topic, like *Query 3*, which is dedicated to calculating the standard variance and exhibits distinct mathematical characteristics, DDASR successfully identifies the correct APIs in its second result. However, it should be noted that the sequence of API calls in this instance is not in the correct order. There are two cases whose query or correct API sequence in the ground truth contains an operation on "*String*" in Table 11. The APIs about "*String*" account for a relatively large portion of the dataset. For queries that involve only processing of *String*, DDASR can generate accurate API sequences. String is only a class used indirectly, like *Query 2*. The most critical APIs are not recommended due to the objective factor of popularity differences.

(4) *DDASR is capable of recommending tail APIs that are useful for developers*. As shown in Table 10, among the tail API dataset, over 77% of the results generated by DDASR were considered practically valuable (whose rating score is greater than or equal to 1) by six developers. The proportion is dramatically higher than those generated by other baselines. This demonstrates that DDASR can effectively recommend useful tail APIs to developers. In Table 11, *Queries 8–10* contain tail APIs in their ground truths. For instance, for *Query 8*, "*DTMNodeProxy.getElementsByTagNameNS*" is a tail API, and DDASR generates "*DTMNodeProxy.getElementsByTagName*," another tail API. Although they are not the same API, they perform similar functions and are clustered into the same cluster during data pre-processing, with "*DTMNodeProxy.getElementsByTagName*" as the cluster center. All six developers think that the API sequences generated for *Query 8* can provide substantial help in resolving the query (rating score equal to 1), demonstrating that developers can derive inspiration from the results generated by DDASR. For *Query 9* and *Query 10*, DDASR successfully hit the tail APIs.

Table 12. Evaluation Results of Different Description Similarity Algorithms of DDASR on the Python Dataset

Similarity Algorithms	$BLEU_O$	$ROUGE_O$	MAP_O	$NDCG_O$	Coverage	Time Cost (ms)
Our Algorithm	0.4858	0.6722	0.5672	0.3677	0.2247	**0.0047**
Cosine Similarity (Word2vec)	0.4805	0.6685	0.5611	0.3624	**0.2249**	0.0169
Cosine Similarity (CodeBERT)	**0.4901**	**0.6782**	**0.5704**	**0.3685**	0.2218	0.0527

The bold text represents the best performance.

Table 13. Evaluation Results on Accuracy of DDASR on the Python Dataset When the Weights of Name Similarity and Description Similarity Vary

α	β	$BLEU_O$	$ROUGE_O$	MAP_O	$NDCG_O$
0	1	0.4799	0.6591	0.5601	0.3622
0.2	0.8	0.4852	0.6668	0.5642	0.3650
0.4	0.6	0.4845	0.6667	0.5624	0.3641
0.6	0.4	0.4856	0.6708	0.5661	0.3673
0.8	0.2	**0.4858**	**0.6722**	**0.5672**	**0.3677**
1	0	0.4852	0.6715	0.5670	0.3671

The bold text represents the best performance.

Answer to RQ3: The human evaluation, conducted by six developers, reveals that the API sequences recommended by DDASR are perceived as useful. Specifically, the developers found it advantageous to have functionally similar cluster centers suggested for tail APIs within sequences that contain tail APIs.

4.6.4 RQ4. How Do the Different Functional Similarity Algorithms and the Parameter Weights in the Similarity Algorithms Affect the Performance of DDASR?

Motivation. We calculate the functional similarity between tail APIs using name similarity and description similarity. In this section, we investigate the performance of DDASR when different similarity algorithms are utilized. Additionally, we study the impact of the weights of name similarity and description similarity on the performance of DDASR.

Approach. We conduct experiments comparing the description similarity calculation algorithm used in DDASR with semantic-based similarity algorithms: one based on Word2vec and the other based on CodeBERT, to assess their impact on the performance of DDASR on the Python dataset. Additionally, we vary the weights of name similarity and description similarity from 0 to 1, in increments of 0.2, and evaluate the performance of DDASR on the Python dataset.

Results. Table 12 shows the evaluation results of different description similarity algorithms of DDASR on the Python dataset. The time cost in the table refers to the average time required to calculate the description similarity between two tail APIs. Table 13 shows the evaluation results on the accuracy of DDASR on the Python dataset when the weights of name similarity and description similarity vary. The values of α and β represent the weights of description similarity and name similarity, respectively, when calculating the functional similarity between tail APIs.

(1) *Different description similarity algorithms do not significantly impact the performance of DDASR; however, the syntactic similarity algorithm we use has an advantage in terms of time cost*. As shown in Table 12, the syntactic similarity algorithm used by DDASR requires the

least computation time, while the cosine similarity algorithm based on CodeBERT takes the longest. In terms of accuracy, the cosine similarity algorithm based on Word2vec presents poor accuracy, while the one based on CodeBERT achieves the best, though the difference is not significant. For diversity, the cosine similarity algorithm based on Word2vec achieves the best coverage, while the one based on CodeBERT performs the worst, and the differences among the three similarity algorithms are also minimal. Compared to the two cosine similarity algorithms, our syntactic similarity algorithm only slightly decreases by a maximum of 0.88%, 0.88%, 0.56%, and 0.22% on the accuracy metrics $BLEU_O$, $ROUGE_O$, MAP_O, and $NDCG_O$, respectively, and by a maximum of 0.09% on the diversity metric. At the same time, it reduces the computation time by at least 72.19%. Given the large number of tail APIs in our task, calculating the description similarity between all pairs of tail APIs requires substantial computational resources. Therefore, balancing both time efficiency and performance in terms of accuracy and diversity, we choose syntactic similarity to calculate the description similarity between tail APIs.

(2) *Descriptions are more important than names when calculating the functional similarity between tail APIs*. As shown in Table 13, DDASR performs the best when $\alpha = 0.8$ and $\beta = 0.2$. However, the weights of description similarity and name similarity do not significantly affect the accuracy performance of DDASR.

Answer to RQ4: Using different similarity algorithms to calculate the description similarity between tail APIs has little impact on the performance of DDASR, but the semantic similarity algorithm we use can significantly reduce time consumption. In the functional similarity between tail APIs, description similarity is more important than name similarity.

5 Threats to Validity

This section discusses potential threats to the validity of our research and experimental design.

Internal Validity. Threats to internal validity, primarily concerning experiment bias and errors, manifest in aspects such as dataset construction and baseline replication [15, 48]. We correct the mismatched parentheses in the Java dataset, which could potentially introduce errors, such as the occurrence of redundant parentheses. To mitigate this threat, our developers conduct a manual inspection, and we have also made the processed dataset publicly available. Moreover, the popularity difference of APIs exists objectively, which is not considered in the original open source Java dataset. To mitigate this threat, we employ the hierarchical sampling method to build a diverse dataset with the same distribution of *query-APIseq* pairs containing tail APIs in the test set as in the training set. Additionally, we conduct multiple checks to ensure that questions in the test set are not included in the training set. Another threat to internal validity, the implementation of baseline approaches, is discussed from the different categories of baselines. For DeepAPI, CodeBERT, and CodeTrans, we directly utilize their open source code or models. Nevertheless, we modify BIKER to make it applicable to our study. For DGAS, we replicate the algorithm and parameters as described in the article. Despite this, there is somewhat of a threat to implementing BIKER, DGAS, and DDASR. To mitigate it, we have conducted multiple code reviews by recruiting experienced programmers and made the source code available on GitHub.

External Validity. External validity refers to the extent to which the results can be generalized beyond the scope of the study. We explore the threats to external validity, which pertain to the generalizability of DDASR. The pairs of query and API sequences in the datasets we use are based on Java and Python. Although the evaluation of two programming languages has validated the generalizability of DDASR to some extent, there is still the possibility that our model may not work

well with other libraries or programming languages. To further validate the model, we will collect more data from other libraries and programming languages for training in the future. Moreover, Gu et al. [21] mined the open source Java dataset by extracting projects from GitHub with more than one star. The constraint indeed introduces some unreliable code. In our evaluation, to maintain a fair comparison with the baselines, we do not modify this restriction. However, to mitigate this threat, we introduce a Python dataset mined by extracting projects from GitHub with more than five stars.

Construct Validity. Threats to construct validity relate to the suitability of our evaluation measures. We use BLEU, a widely used metric in the translation task, to evaluate the accuracy of API sequences, because generating API sequences from a query can be analogous to translating. MAP and NDCG are the classical accuracy evaluation measures in recommender systems, which are also widely used in the software engineering field. Coverage is widely used in diversity recommendation as an evaluation metric for the proportion of recommended items. In addition, to assess the impact of the pseudo ground truth on evaluating authenticity, we introduce a restored veritable accuracy metric to mitigate bias. As a result of different coding habits, developers have different ways of implementing requirements when completing development tasks. However, there is only one correct way to solve the query in the ground truth of the dataset. We utilize a user study to mitigate this threat through manual evaluation by developers with programming experience.

Conclusion Validity. Conclusion validity concerns how much the experiment setting influences the observed result. The baseline approaches split APIs into fragments for data pre-processing and experimental evaluation, while our study regards a complete API as a basic unit for the two stages. This difference presents a potential risk to conclusion validity. To address this, we conduct an ablation experiment with DASR, a variant of our approach, assessing the impact of training solely with aggregated API fragments.

6 Conclusion and Future Work

In this article, we propose DDASR to recommend API sequences automatically for developer queries. Our study highlights the importance of incorporating diversity in API sequence recommendation, particularly with tail APIs. DDASR can recommend functionally similar tail APIs, helping to mitigate the long-tail effect. For programming tasks, recommending APIs with similar functionality can provide heuristic assistance to developers. We achieve this by clustering tail APIs based on function and substituting them with cluster centers to create a pseudo ground truth, followed by recommendations based on Seq2Seq and LTR techniques. Due to the outstanding performance of LLMs in natural language processing tasks, we also leverage widely adopted LLMs for learning query representations. The evaluation with state-of-the-art baselines on the original Java dataset confirms DDASR's accuracy. Experiments on the diverse Java dataset and the Python dataset show that DDASR can achieve the best diversity without significantly reducing accuracy.

In DDASR, we utilize lightweight syntactic similarity to calculate the functional similarity between tail APIs. Moving forward, we plan to explore the performance of other similarity algorithms in clustering tail APIs. Additionally, DDASR provides developers with cluster centers of similar functionality among tail APIs due to the construction of the pseudo ground truth. In the future, we will also consider how to achieve more precise recommendations of tail APIs. Considering the personalized coding styles of developers, multiple API combinations may fulfill the same requirement, suggesting that the API sequence in the ground truth is not the only correct solution. In the future, we hope to satisfy personalized and multi-objective API sequence recommendations from various aspects of dataset construction, model improvement, and innovation in evaluation metrics. We also plan to integrate a broader range of refined tail APIs into the sequences to further improve DDASR's performance. Furthermore, we intend to apply DDASR across more programming languages and a

wider range of requirement domains. LLMs perform well in automated code generation tasks, while they tend to write code rather than apply existing APIs [69]. Focusing on the importance of API recommendation in automated code generation [41] to enhance the readability and maintainability of code generated by LLMs will be a future research direction for us.

References

[1] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911. DOI: https://doi.org/10.1109/TKDE.2011.15

[2] Gediminas Adomavicius and YoungOk Kwon. 2014. Optimization-based approaches for maximizing aggregate recommendation diversity. *INFORMS Journal on Computing* 26, 2 (2014), 351–369. DOI: https://doi.org/10.1287/ijoc.2013.0570

[3] Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Unified pre-training for program understanding and generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '21)*, 2655–2668.

[4] Chris Anderson. 2012. The long tail. In *The Social Media Reader*. Michael Mandiberg (Ed.), New York University Press, 137–152. DOI: https://doi.org/10.18574/nyu/9780814763025.003.0014

[5] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural probabilistic language models. In *Innovations in Machine Learning: Theory and Applications*. Holmes E. Dawn and Jain C. Lakhmi (Eds.), Springer, Berlin, 137–186. DOI: https://doi.org/10.1007/3-540-33486-6_6

[6] Steven Bird. 2006. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL Interactive Presentation Sessions (ACL '06)*, 69–72. DOI: https://doi.org/10.3115/1225403.1225421

[7] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, 335–336. DOI: https://doi.org/10.1145/290941.291025

[8] Wing-Kwan Chan, Hong Cheng, and David Lo. 2012. Searching connected API subgraph via text phrases. In *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering (SIGSOFT FSE '12)*, 1–11. DOI: https://doi.org/10.1145/2393596.2393606

[9] Chi Chen, Xin Peng, Bihuan Chen, Jun Sun, Zhenchang Xing, Xin Wang, and Wenyun Zhao. 2022. "More than deep learning": Post-processing for API sequence recommendation. *Empirical Software Engineering* 27 (2022), 1–32. DOI: https://doi.org/10.1007/s10664-021-10040-2

[10] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy MAP inference for determinantal point process to improve recommendation diversity. arXiv:1709.05135. Retrieved from https://arxiv.org/abs/1709.05135

[11] Yujia Chen, Cuiyun Gao, Xiaoxue Ren, Yun Peng, Xin Xia, and Michael R. Lyu. 2023. API usage recommendation via multi-view heterogeneous graph representation learning. *IEEE Transactions on Software Engineering* 49, 5 (2023), 3289–3304. DOI: https://doi.org/10.1109/TSE.2023.3252259

[12] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to recommend accurate and diverse items. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*, 183–192. DOI: https://doi.org/10.1145/3038912.3052585

[13] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT'S capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys '23)*, 1126–1132. DOI: https://doi.org/10.1145/3604915.3610646

[14] Ahmed Elnaggar, Wei Ding, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Silvia Severini, Florian Matthes, and Burkhard Rost. 2021. CodeTrans: Towards cracking the language of silicon's code through self-supervised deep learning and high performance computing. arXiv:2104.02443. Retrieved from https://arxiv.org/abs/2104.02443

[15] Robert Feldt and Ana Magazinius. 2010. Validity threats in empirical software engineering research - An initial survey. In *Proceedings of the 22nd International Conference on Software Engineering Knowledge Engineering (SEKE '10)*, 374–379.

[16] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, et al. 2020. CodeBERT: A pre-trained model for programming and natural languages. In *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP Findings '20)*, 1536–1547. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.139

[17] Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. arXiv:2305.07961. Retrieved from https://arxiv.org/abs/2305.07961

[18] Junchen Fu, Fajie Yuan, Yu Song, Zheng Yuan, Mingyue Cheng, Shenghui Cheng, Jiaqi Zhang, Jie Wang, and Yunzhu Pan. 2023. Exploring adapter-based transfer learning for recommender systems: Empirical studies and practical insights. arXiv:2305.15036. Retrieved from https://arxiv.org/abs/2305.15036

[19] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Deutschland GmbH.

[20] Marko Gasparic and Andrea Janes. 2016. What recommendation systems for software engineering recommend: A systematic literature review. *Journal of Systems and Software* 113 (2016), 101–113. DOI: https://doi.org/10.1016/j.jss.2015.11.036

[21] Xiaodong Gu, Hongyu Zhang, Dongmei Zhang, and Sunghun Kim. 2016. Deep API learning. In *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (SIGSOFT FSE '16)*, 631–642. DOI: https://doi.org/10.1145/2950290.2950334

[22] Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. 2022. UniXcoder: Unified cross-modal pre-training for code representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL '22)*, 7212–7225. DOI: https://doi.org/10.18653/v1/2022.acl-long.499

[23] Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, et al. 2021. GraphCode{BERT}: Pre-training code representations with data flow. In *Proceedings of the International Conference on Learning Representations (ICLR '21)*.

[24] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. arXiv:1511.06939. Retrieved from https://arxiv.org/abs/1511.06939

[25] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2023. Large language models are zero-shot rankers for recommender systems. arXiv:2305.08845. Retrieved from https://arxiv.org/abs/2305.08845

[26] Qiao Huang, Xin Xia, Zhenchang Xing, David Lo, and Xinyu Wang. 2018. API method recommendation without worrying about the task-API knowledge gap. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE '18)*, 293–304. DOI: https://doi.org/10.1145/3238147.3238191

[27] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2020. CodeSearchNet challenge: Evaluating the state of semantic code search. arXiv:1909.09436. Retrieved from https://arxiv.org/abs/1909.09436

[28] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. arXiv:1412.2007. Retrieved from https://arxiv.org/abs/1412.2007

[29] Zhengbao Jiang, Zhicheng Dou, Wayne Xin Zhao, Jian-Yun Nie, Ming Yue, and Ji-Rong Wen. 2018. Supervised search result diversification via subtopic attention. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1971–1984. DOI: https://doi.org/10.1109/TKDE.2018.2810873

[30] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems* 7, 1 (2016), 1–42. DOI: https://doi.org/10.1145/2926720

[31] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM '18)*, 197–206. DOI: https://doi.org/10.1109/ICDM.2018.000356

[32] Yejin Kim, Kwangseob Kim, Chanyoung Park, and Hwanjo Yu. 2019. Sequential and diverse recommendation with long tail. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI '19)*, 2740–2746. DOI: https://doi.org/10.24963/ijcai.2019/380

[33] Amy J. Ko, Brad A. Myers, and Htet Htet Aung. 2004. Six learning barriers in end-user programming systems. In *Proceedings of the IEEE Symposium on Visual Languages - Human Centric Computing (VL/HCC '04)*, 199–206. DOI: https://doi.org/10.1109/VLHCC.2004.47

[34] Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the Machine Translation: From Real Users to Research (AMTA '04)*, 115–124. DOI: https://doi.org/10.1007/978-3-540-30194-3_13

[35] Xianglong Kong, Weina Han, Li Liao, and Bixin Li. 2020. An analysis of correctness for API recommendation: Are the unmatched results useless? *Science China Information Sciences* 63 (2020), 1–15. DOI: https://doi.org/10.1007/s11432-019-2929-9

[36] Hyokmin Kwon, Jaeho Han, and Kyungsik Han. 2020. ART (attractive recommendation tailor): How the diversity of product recommendations affects customer purchase preference in fashion industry? In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, 2573–2580. DOI: https://doi.org/10.1145/3340531.3412687

[37] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444. DOI: https://doi.org/10.1038/nature14539

[38] Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10 (1966), 707–710.

[39] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*, 1419–1428. DOI: https://doi.org/10.1145/3132847.3132926

[40] Shuang Li, Yuezhi Zhou, Di Zhang, Yaoxue Zhang, and Xiang Lan. 2017. Learning to diversify recommendations based on matrix factorization. In *Proceedings of the IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech '17)*, 68–74. DOI: https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.26

[41] Dianshu Liao, Shidong Pan, Xiaoyu Sun, Xiaoxue Ren, Qing Huang, Zhenchang Xing, Huan Jin, and Qinying Li. 2024. A3-CodGen : A repository-level code generation framework for code reuse with local-aware, global-aware, and third-party-library-aware. *IEEE Transactions on Software Engineering* 01 (2024), 1–16. Retrieved from https://doi.ieeecomputersociety.org/10.1109/TSE.2024.3486195

[42] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, et al. 2023. How can recommender systems benefit from large language models: A survey. arXiv:2306.05817. Retrieved from https://arxiv.org/abs/2306.05817

[43] Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a good recommender? A preliminary study. arXiv:2304.10149. Retrieved from https://arxiv.org/abs/2304.10149

[44] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. arXiv:1508.04025. Retrieved from https://arxiv.org/abs/1508.04025

[45] Homan Ma, Robert Amor, and Ewan Tempero. 2006. Usage patterns of the Java standard API. In *Proceedings of the 13th Asia Pacific Software Engineering Conference (APSEC '06)*, 342–352. DOI: https://doi.org/10.1109/APSEC.2006.60

[46] James Martin and Jin L. C. Guo. 2022. Deep API learning revisited. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension (ICPC '22)*, 321–330. DOI: https://doi.org/10.1145/3524610.3527872

[47] Collin McMillan, Mark Grechanik, Denys Poshyvanyk, Qing Xie, and Chen Fu. 2011. Portfolio: Finding relevant functions and their usage. In *Proceedings of the 33rd International Conference on Software Engineering (ICSE '11)*, 111–120. DOI: https://doi.org/10.1145/1985793.1985809

[48] Nasser Mustafa, Yvan Labiche, and Dave Towey. 2019. Mitigating threats to validity in empirical software engineering: A traceability case study. In *Proceedings of the IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC '19)*, 324–329. DOI: https://doi.org/10.1109/COMPSAC.2019.10227

[49] Phuong T. Nguyen, Juri Di Rocco, Davide Di Ruscio, and Massimiliano Di Penta. 2020. CrossRec: Supporting software developers by recommending third-party libraries. *Journal of Systems and Software* 161 (2020), 110460. DOI: https://doi.org/10.1016/j.jss.2019.110460

[50] Haoran Niu, Iman Keivanloo, and Ying Zou. 2017. Learning to rank code examples for code search engines. *Empirical Software Engineering* 22 (2017), 259–291. DOI: https://doi.org/10.1007/s10664-015-9421-5

[51] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2024. GPT-4 technical report. arXiv:2303.08774. Retrieved from https://arxiv.org/abs/2303.08774

[52] Ali Ouni, Raula Gaikovina Kula, Marouane Kessentini, Takashi Ishio, Daniel M. German, and Katsuro Inoue. 2017. Search-based software library recommendation using multi-objective optimization. *Information and Software Technology* 83 (2017), 55–75. DOI: https://doi.org/10.1016/j.infsof.2016.11.007

[53] Yoon-Joo Park. 2012. The adaptive clustering method for the long tail problem of recommender systems. *IEEE Transactions on Knowledge and Data Engineering* 25, 8 (2012), 1904–1915. DOI: https://doi.org/10.1109/TKDE.2012.119

[54] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)*, 11–18. DOI: https://doi.org/10.1145/1454008.1454012

[55] Yun Peng, Shuqing Li, Wenwei Gu, Yichen Li, Wenxuan Wang, Cuiyun Gao, and Michael R. Lyu. 2023. Revisiting, benchmarking and exploring API recommendation: How far are we? *IEEE Transactions on Software Engineering* 49, 4 (2023), 1876–1897. DOI: https://doi.org/10.1109/TSE.2022.3197063

[56] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys '17)*, 130–137. DOI: https://doi.org/10.1145/3109859.3109896

[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[58] Mukund Raghothaman, Yi Wei, and Youssef Hamadi. 2016. SWIM: Synthesizing what I mean - Code search and idiomatic snippet synthesis. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*, 357–367. DOI: https://doi.org/10.1145/2884781.2884808

[59] Mohammad Masudur Rahman, Chanchal K. Roy, and David Lo. 2016. RACK: Automatic API recommendation using crowdsourced knowledge. In *Proceedings of the IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER '16)*, 349–359. DOI: https://doi.org/10.1109/SANER.2016.80

[60] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, 811–820. DOI: https://doi.org/10.1145/1772690.17727738

[61] Martin Robillard, Robert Walker, and Thomas Zimmermann. 2010. Recommendation systems for software engineering. *IEEE Software* 27, 4 (2010), 80–86. DOI: https://doi.org/10.1109/MS.2009.161

[62] Hinrich Schütze, Christopher D. Manning, and Prabhakar Raghavan. 2008. *Introduction to Information Retrieval*, Vol. 39. Cambridge University Press, Cambridge.

[63] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. arXiv:1508.07909. Retrieved from https://arxiv.org/abs/1508.07909

[64] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? Investigating large language models as re-ranking agents. arXiv:2304.09542. Retrieved from https://arxiv.org/abs/2304.09542

[65] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM '18)*, 565–573. DOI: https://doi.org/10.1145/3159652.3159656

[66] Ferdian Thung, David Lo, and Julia Lawall. 2013. Automated library recommendation. In *Proceedings of the 20th Working conference on reverse engineering (WCRE '13)*, 182–191. DOI: https://doi.org/10.1109/WCRE.2013.6671293

[67] Yuan Tian, Ferdian Thung, Abhishek Sharma, and David Lo. 2017. APIBot: Question answering bot for API documentation. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE '17)*, 153–158. DOI: https://doi.org/10.1109/ASE.2017.8115628

[68] Von Luxburg Ulrike. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 395–416. DOI: https://doi.org/10.1007/s11222-007-9033-z

[69] Yanlin Wang, Tianyue Jiang, Mingwei Liu, Jiachi Chen, and Zibin Zheng. 2024. Beyond functional correctness: Investigating coding style inconsistencies in large language models. arXiv:2407.00456. Retrieved from https://arxiv.org/abs/2407.00456

[70] Yue Wang, Weishi Wang, Shafiq Joty, and Steven C. H. Hoi. 2021. CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*, 8696–8708. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.685

[71] Jacek Wasilewski and Neil Hurley. 2016. Incorporating diversity in a learning to rank recommender system. In *Proceedings of the 29th International Flairs Conference (FLAIRS '16)*, 572–578.

[72] Hongwei Wei, Xiaohong Su, Weining Zheng, and Wenxin Tao. 2023. Documentation-guided API sequence search without worrying about the text-API semantic gap. In *Proceedings of the IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER '23)*, 343–354. DOI: https://doi.org/10.1109/SANER56733.2023.00040

[73] Moshi Wei, Nima Shiri Harzevili, Yuchao Huang, Junjie Wang, and Song Wang. 2022. CLEAR: Contrastive learning for API recommendation. In *Proceedings of the 44th International Conference on Software Engineering (ICSE '22)*, 376–387. DOI: https://doi.org/10.1145/3510003.3510159

[74] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical diversified recommendations on YouTube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, 2165–2173. DOI: https://doi.org/10.1145/3269206.3272018

[75] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, Alexander J. Smola, and How Jing. 2017. Recurrent recommender networks. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM '17)*, 495–503. DOI: https://doi.org/10.1145/3018661.3018689

[76] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Bowei He, Bhaskar Mitra, and Xue Liu. 2023. Result diversification in search and recommendation: A survey. arXiv:2212.14464. Retrieved from https://arxiv.org/abs/2212.14464

[77] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, 1192–1199. DOI: https://doi.org/10.1145/1390156.1390306

[78] Hongzhi Yin, Bin Cui, Jing Li, Junjie Yao, and Chen Chen. 2012. Challenging the long tail recommendation. *Proceedings of the VLDB Endowment* 5, 9 (2012), 896–907. DOI: https://doi.org/10.14778/2311906.2311916

[79] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M. Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*, 582–590. DOI: https://doi.org/10.1145/3289600.3290975

[80] Jingxuan Zhang, He Jiang, Zhilei Ren, and Xin Chen. 2018. Recommending APIs for API related questions in stack overflow. *IEEE Access* 6 (2018), 6205–6219. DOI: https://doi.org/10.1109/ACCESS.2017.2777845

[81] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-news matching BERT for news recommendation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI '21)*, 3356–3362. DOI: https://doi.org/10.24963/ijcai.2021/462

[82] Shuai Zhang, Yi Tay, Lina Yao, and Aixin Sun. 2018. Next item recommendation with self-attention. arXiv:1808.06414. Retrieved from https://arxiv.org/abs/1808.06414

[83] Yuhui Zhang, Hao Ding, Zeren Shui, Yifei Ma, James Zou, Anoop Deoras, and Hao Wang. 2021. Language models as recommender systems: Evaluations and limitations. In *Proceedings of the I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop*.

[84] Yu Zheng, Chen Gao, Liang Chen, Depeng Jin, and Yong Li. 2021. DGCN: Diversified recommendation with graph convolutional networks. In *Proceedings of the Web Conference (WWW '21)*, 401–412. DOI: https://doi.org/10.1145/3442381.3449835

[85] Hao Zhong and Hong Mei. 2019. An empirical study on API Usages. *IEEE Transactions on Software Engineering* 45, 4 (2019), 319–334. DOI: https://doi.org/10.1109/TSE.2017.2782280

[86] Hao Zhong, Tao Xie, Lu Zhang, Jian Pei, and Hong Mei. 2009. MAPO: Mining and recommending API usage patterns. In *Proceedings of the 23rd European Conference Object-Oriented Programming (ECOOP '09)*, 318–343. DOI: https://doi.org/10.1007/978-3-642-03013-0_15

[87] Jianghong Zhou, Eugene Agichtein, and Surya Kallumadi. 2020. Diversifying multi-aspect search results using Simpson's diversity Index. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, 2345–2348. DOI: https://doi.org/10.1145/3340531.3412163

[88] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web (WWW '05)*, 22–32. DOI: https://doi.org/10.1145/1060745.1060754

Received 10 January 2024; revised 11 December 2024; accepted 17 December 2024