**A. Validation Dataset**

To validate the accuracy and robustness of the proposed method for knowledge extraction, we manually sampled 30 typical Google Earth Engine (GEE) workflow scripts from the GEE script collection, considering various dimensions such as length, modeling target, and functionality. We constructed GEE knowledge templates using these GEE workflow scripts to perform validation experiments. Besides, based on expert knowledge, we annotated the ground truth of temporal range, spatial scope, data source, and thematic task dimensions for these samples. These ground truth will be compared with the descriptive knowledge extracted by GPT models, specifically GPT-3.5-turbo and GPT-4, which are employed in our experiments.

On the other hand, to assess the stability and reliability of procedural knowledge extraction, we obtained directed acyclic graphs (DAGs) for all samples in the GEE environment, which include the pairwise connections between APIs in each sample. We manually filtered and constructed ground truth tables based on expert knowledge, where each row in the ground truth table represents a connection between two APIs. Since the API connections are consolidated in the ServiceNode and Relationship sub-nodes under the Process node in the GEE knowledge template, the API connections in the knowledge template will be used as predictions and compared with the ground truth tables based on fuzzy matching. The evaluation will use accuracy, recall, and F1 score as metrics.

**B. Evaluation Methods and Results Analysis**

To evaluate the stability and completeness of the GPT models’ extraction, and to test whether the uncertainty characteristics of them affect the extraction results, we set the hyperparameter temperature to 0.2, which is considered suitable for inference tasks. We ran 5 iterations of the GPT-3.5-turbo and GPT-4 models and conected their responses. Specifically, we analyzed the results generated in each round of the GPT model and evaluated the stability by comparing the frequency of the most common values with the number of iterations. For instance, if a thematic task sample appeared as “Land use/land cover” in three out of five iterations, the stability value for that sample would be 0.6. Finally, the average stability values for all samples are calculated to assess the extraction stability of the GPT models, as shown in Table 1.

Table 1. Stability of descriptive knowledge extraction results on average

| **GPT Model Name** | **Temporal Range** | **Spatial Scope** | **Data Source** | **Thematic Task** |
| --- | --- | --- | --- | --- |
| GPT-3.5 Turbo | 0.99 | 0.97 | **1.0** | 0.94 |
| GPT-4 | **1.0** | 0.99 | **1.0** | 0.95 |

The results of the stability evaluation indicate that under the constraint of the hyperparameter temperature (set to 0.2), all GPT models can produce stable results across all dimensions, with almost identical responses over the 5 iterations. This suggests that in the knowledge extraction tasks presented in this study, the generated content is not significantly affected by the uncertainty of the GPT models, which also indicates the consistency of the outputs. Only a few samples showed instability in their evaluation results (for detailed records, see the stability\_scores\_with\_avg.csv file). However, the stability of the generated results does not necessarily imply their accuracy. Therefore, the most frequent value from the results will be used as the prediction to compare with the ground truth to evaluate the accuracy of these GPT models.

We compared the accuracy of descriptive knowledge extraction between the GPT-3.5-turbo and GPT-4 models. Specifically, if the prediction matches the corresponding ground truth value, the result will be labeled as 1, indicating its correctness; otherwise, it will be marked as 0. By summarizing all results, we obtained the final accuracy evaluation for each dimension. As shown in Fig. 1, the results demonstrate that the GPT models are capable of summarizing GEE workflow scripts and, leveraging their strong reasoning abilities, generate values that meet the requirements according to the prompt engineering (see Section 3.2 in the manuscript). These values are either in a user-defined structured format or from a provided list of values (GCMD keywords).

图表, 条形图

描述已自动生成

Fig. 1. Comparison of descriptive knowledge extraction results between GPT-3.5 Turbo and GPT-4 models

The statistical results indicate that the GPT-4 model outperforms the GPT-3.5 Turbo model in every knowledge extraction dimension. Both GPT models perform best in the temporal dimension, as they are able to sensitively detect whether each sample includes a time range or specific start and end times. In the Data Source and Geospatial Scope dimensions, the GPT-4 model shows a significant improvement over GPT-3.5 Turbo. For example, in the Data Source dimension, GPT-4 achieves an accuracy of 0.9, which means the model can almost perfectly identify the data content used in the provided samples. However, we argued that in the current accuracy metric, we only consider the extraction result correct if the data sources included in the prediction are entirely consistent with the ground truth. In fact, several of the data source information extracted by GPT-3.5 Turbo is incomplete. For instance, if a sample contains 5 data sources and GPT-3.5 Turbo identifies 3 of them, the result will still be marked as 0, even though this does not mean that GPT-3.5 Turbo has completely failed in data source extraction. Therefore, the actual difference in capability between the tested GPT models in the data source dimension is likely to be smaller than presented. Likewise, the performance gap between GPT-3.5 Turbo and GPT-4 models in thematic tasks is relatively small. In this experiment, we used only one level of GCMD keywords as alternative values and required the GPT models to select just one of them. If we allowed the selection of multiple thematic tasks, which is reasonable in practical geographic analysis models, the accuracy results would differ. Nevertheless, the current results still demonstrate the GPT models’ abilities to extract knowledge in the thematic task dimension. Therefore, we will conduct further evaluations of knowledge extraction under more complex conditions for thematic tasks in the future work.

Finally, we evaluated the accuracy, recall, and F1 score of the procedural knowledge extraction for each script by comparing the API pairs in the predictions with those in the ground truth tables. We calculated the average of these metrics to obtain the final evaluation results, as shown in Table 2.

Table 2. Accuracy, recall and F1 score for the procedural knowledge extraction results

|  |  |  |  |
| --- | --- | --- | --- |
| **Script Name** | **Precision** | **Recall** | **F1 Score** |
| script\_320000 | 1.0 | 1.0 | 1.0 |
| script\_320001 | 1.0 | 0.938 | 0.968 |
| script\_320002 | 1.0 | 0.957 | 0.978 |
| script\_320003 | 1.0 | 1.0 | 1.0 |
| script\_320004 | 1.0 | 0.882 | 0.938 |
| script\_320005 | 1.0 | 0.875 | 0.933 |
| script\_320006 | 1.0 | 0.571 | 0.727 |
| script\_320007 | 1.0 | 1.0 | 1.0 |
| script\_320008 | 0.909 | 0.909 | 0.909 |
| script\_320009 | 0.692 | 0.562 | 0.621 |
| script\_320010 | 1.0 | 0.667 | 0.8 |
| script\_320011 | 0.846 | 0.786 | 0.815 |
| script\_320012 | 1.0 | 1.0 | 1.0 |
| script\_320013 | 1.0 | 1.0 | 1.0 |
| script\_320014 | 1.0 | 0.8 | 0.889 |
| script\_320015 | 1.0 | 1.0 | 1.0 |
| script\_320016 | 1.0 | 1.0 | 1.0 |
| script\_320017 | 1.0 | 1.0 | 1.0 |
| script\_320018 | 1.0 | 0.6 | 0.75 |
| script\_320019 | 1.0 | 0.688 | 0.815 |
| script\_320020 | 0.958 | 0.92 | 0.939 |
| script\_320021 | 1.0 | 0.75 | 0.857 |
| script\_320022 | 1.0 | 1.0 | 1.0 |
| script\_320023 | 1.0 | 1.0 | 1.0 |
| script\_320024 | 1.0 | 0.944 | 0.971 |
| script\_320025 | 1.0 | 1.0 | 1.0 |
| script\_320026 | 1.0 | 0.833 | 0.909 |
| script\_320027 | 1.0 | 0.789 | 0.882 |
| script\_320028 | 1.0 | 0.9 | 0.947 |
| script\_320029 | 1.0 | 1.0 | 1.0 |
| **Average** | 0.98 | 0.879 | 0.922 |

According to results in Table 2, it can be observed that the AST-based modeling process extraction method (See Section 3.3 in the manuscript) was able to identify the majority of API pairs in most samples, with an accuracy of 1 in many cases. This indicates that every extracted API pair could be matched with an entry in the ground truth. In some cases, the lower accuracy is due to the reason that the proposed algorithm incorrectly extracted methods that do not exist in the GEE. On the other hand, the average recall was slightly lower, indicating that there is still space for improving the algorithm to identify and extract the relationships between APIs in certain situations. Nevertheless, the overall F1 score, which combines accuracy and recall, was 0.922, demonstrating the great performance of the AST-based approach on the samples. Therefore, the extraction results for both descriptive knowledge and procedural knowledge validate the feasibility of our proposed method, resulting in high-quality of the GEE knowledge templates.