GDCF: A Generalizable Digital Cognition Framework Incorporating Teacher Model Generated Pseudo-label Verification into Instruction Tuning

Anonymous ACL submission

Abstract

To address the semantic gap between specialized terminology in cultural heritage and everyday public language, this paper innovatively proposes the Generalizable Digital Cognition Framework (GDCF), focusing on overcoming cross-domain semantic alignment challenges in low-resource scenarios. By leveraging a teacher-student model architecture and instruction tuning techniques, GDCF achieves accurate mapping from everyday language to domain-specific vocabulary in a few-shot setting with only 100 annotated samples. The teacher model generates initial pseudo-labels, while a dynamic label masking strategy guides the smaller student model through instruction tuning, enabling it to achieve performance comparable to the teacher model. Remarkably, when both teacher and student models use the same parameter size, the student model can even outperform the teacher model. Experiments show that this method achieves a keyword extraction accuracy of 0.39 on a cultural heritage review dataset, marking a 73% improvement over the baseline LLM. More significantly, this framework pioneers a 3D visualization space that integrates semantic vectors with cognitive dynamics, uncovering deep semantic relationships between public discourse and professional terminology. Its modular design has been successfully validated for transferability in architectural heritage conservation assessments, providing a scalable benchmark paradigm for interdisciplinary digital humanities research.

1 Introduction

Cultural heritage, a key carrier of human civilization, influences how culture is shared. Professionals use structured terminology to describe architectural heritage, while the public relies on informal expressions, creating a gap that limits the dissemination of academic insights and complicates the extraction of cultural knowledge from public discourse. Existing cultural heritage databases largely provide one-way information flow, lacking dynamic interaction between expert semantics and public language.

Cross-lingual and cross-domain semantic alignment in cultural heritage faces three challenges: (1) Static mappings—predefined vocabularies fail to adapt to evolving terminology, causing semantic gaps; (2) Low domain adaptability—general-purpose models struggle with specialized data, particularly ancient architectural terms; (3) Unidirectional cognition—current systems translate expert language into public expressions but rarely extract professional semantics from unstructured commentary, failing to complete a "professional–public–professional" feedback loop.

To address these challenges, this paper introduces the Generalizable Digital Cognition Framework (GDCF), which transforms unstructured everyday language into professional terminology through a three-stage process, creating a computable, interactive 3D semantic space for cross-domain research.

Task 1: Domain-Specific Dictionary Construction – Using the cultural heritage domain, which is highly specialized but data-scarce, as a case study, we develop a transferable domain dictionary system. This system systematically integrates specialized vocabulary from cultural and mixed heritage sections of the World Heritage List, establishing the first automatically generated semantic ontology for world cultural heritage, filling a gap in structured knowledge representation in this field.

Task 2: Mapping from Everyday Language to Professional Vocabulary – Using a teacher-student model and instruction tuning, a 0.5B-parameter model is fine-tuned on only 100 samples. This model is then applied to 30,000 Weibo comments on cultural heritage, extracting keywords that map everyday language to professional terminology.

Task 3: 3D Semantic Space Modeling –

Based on the 30,000 Weibo comments, we construct a three-dimensional model that integrates professional semantics and cognitive dynamics. This framework pioneers mathematical modeling and dynamic visualization of the "everyday-to-professional" semantic space in the humanities and social sciences. Its modular design allows rapid transferability to fields such as history and art, providing a scalable benchmark paradigm for interdisciplinary digital research.

2 Related Work

2.1 Instruction Tuning

Instruction tuning refers to the process of adjusting a language model to better follow natural-language instructions. Unlike traditional supervised learning, which requires large task-specific datasets, instruction tuning trains models on a broad range of tasks formulated as instructions. This enables greater adaptability and generalization, allowing models to handle new tasks with minimal or no additional training.

(Ouyang et al., 2022) introduce a fine-tuning approach that incorporates human feedback to improve instruction adherence. Their method includes supervised fine-tuning with human-written examples, training a reward model based on human-labeled response preferences, and applying reinforcement learning with human feedback (RLHF) using Proximal Policy Optimization (PPO). This process effectively aligns model behavior with user expectations, reducing biases and improving response truthfulness. Notably, their results show that a 1.3B parameter InstructGPT model outperforms the much larger 175B GPT-3 in human evaluations. However, challenges such as annotator bias and computational cost highlight the need for more efficient alignment methods.

(Wang et al., 2023) propose SELF-INSTRUCT, an instruction tuning framework that reduces reliance on human-written datasets by having the model generate and refine its own instructions. Through an iterative process, the model generates task instructions, synthesizes input-output pairs, and filters low-quality instructions before fine-tuning. Applied to GPT-3, SELF-INSTRUCT produced 52K instructions, leading to a 33% improvement on the SUPER-NATURALINSTRUCTIONS benchmark, nearly matching OpenAI's InstructGPT-001. While scalable, this method faces challenges such as potential biases and lack of human oversight, suggesting the need for hybrid human-machine validation.

These works demonstrate two key directions in instruction tuning: leveraging human feedback for fine-tuning and automating instruction generation. While both approaches significantly enhance instruction-following capabilities, the trade-offs between human supervision, scalability, and bias mitigation remain open challenges, motivating further research into more efficient, scalable, and aligned instruction tuning methods.

2.2 Keyword Extraction

Keyword extraction is a core task in NLP, identifying key terms from unstructured text to build structured dictionaries. Traditional methods rely on statistical features (e.g., TF-IDF, TextRank) or rule-based templates but struggle with domain adaptability and low-frequency terms.

Recent unsupervised approaches combine statistical features with semantic patterns for cross-domain extraction. YAKE assigns term weights based on position, frequency, and word associations, making it effective in resource-limited settings (Campos et al., 2020). Pretrained models like BERT (Devlin et al., 2019) and FastText (Joulin et al., 2016) enhance dictionary quality by learning semantic representations from large corpora. FastText, similar to CBOW (Mikolov et al., 2013), maps words to vectors, averages them, and feeds them into a classifier using softmax for probability distribution.

(Zhu et al., 2024) propose a pretrained language model (PLM) integrating domain-specific heterogeneous knowledge. By combining unstructured, semi-structured, and structured texts, the model captures richer contextual information. It simultaneously models entity descriptions, titles, and knowledge triples within a shared space, overcoming traditional text-only limitations. An unsupervised pretraining strategy enhances entity and topic knowledge representation, improving performance in complex document analysis tasks.

While NLP advances enable efficient keyword extraction, cultural heritage still lacks specialized dictionaries enriched with domain knowledge and knowledge graph associations. Dictionary construction remains largely manual, with limited machine-assisted applications.

2.3 Lexical Semantic Mapping

Lexical semantic mapping establishes correspondences between domain-specific terms and their generalized expressions while preserving key concepts. Traditional methods rely on manually constructed semantic networks or bilingual dictionaries, which lack scalability. Embedding-based methods address this by projecting words into continuous vector spaces, where cosine similarity and Euclidean distance quantify semantic relationships for automated term alignment.

(Bosc and Vincent, 2018) propose a recursive autoencoder model (CPAE) that leverages dictionary definitions for term embedding. Without large corpora, CPAE encodes semantic similarities and distinctions into a generalized semantic space. For example, hostage ("a prisoner held to ensure terms") is mapped near prisoner while retaining uniqueness, providing an efficient approach to semantic normalization.

(Trifonov et al., 2018) enhance cross-domain interpretability with a sparse sentence embedding method using k-Sparse and Sparsemax constraints. Their approach captures semantic units and quantifies semantic correlations, making it adaptable to lexical alignment tasks.

By integrating structured knowledge (e.g., dictionary definitions) with sparse semantic encoding, these methods improve accuracy and interpretability in cross-domain lexical mapping.

3 Dataset

3.1 Data Design

We constructed two datasets with sharply contrasting characteristics to encompass the broad linguistic contexts of both experts and the general public. Dataset a is a specialized expert-language corpus, characterized by high credibility and rigor, while dataset b contains everyday-language texts related to the World Cultural Heritage sites "the Imperial Palace" and "the Potala Palace", as well as the Chinese cultural heritage site "the Old Summer Palace". In this way, dataset b reflects linguistic features from a non-expert perspective.

3.2 Data Collection

For dataset a, we referenced the UNESCO World Heritage List and excluded natural heritage entries, selecting only the official introductory articles for cultural and mixed heritage sites. A total of 953 articles were included. For dataset b, we used publicly available data from the Chinese social media platform Sina Weibo (whose monthly active users reached 598 million according to its Q4 2023 financial report). We collected all user-generated posts containing the keywords "the Imperial Palace", "the Potala Palace", and "the Old Summer Palace" from June to December 2023. Based on this, we formed three sub-datasets: b1 for the imperial palace, b2 for the Potala Palace, and b3 for the Old Summer Palace, thereby assembling a rich non-expert corpus.

3.3 Data Cleaning

To improve data quality and facilitate model adaptation, we used a range of noise-reduction strategies. First, we removed duplicate sentences to maintain independence across data samples. We then eliminated irrelevant symbols and noise, such as HTML tags, garbled text, advertisements, copyright notices, and navigation content introduced by online sources. We also corrected spelling errors. Lastly, we removed unlabeled multilingual content to keep the corpus consistent in its primary language.

3.4 Data Annotation

First, we performed partial manual annotation on the dataset to facilitate few-shot instruction tuning in the subsequent large-model fine-tuning module. We labeled keywords from 90 professional sentences in dataset A and randomly selected 10 sentences from dataset B, marking them with professional terms. The 100 annotated sentences were then compiled into dataset C.

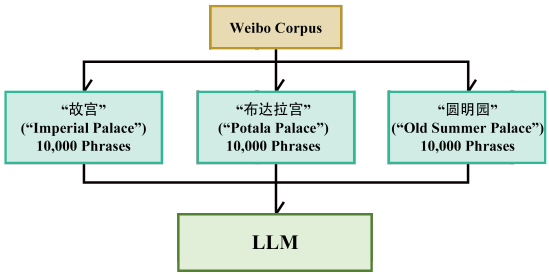

Figure 1: Weibo Corpus Annotation

For these sentences, we annotated all professionally relevant terms, including but not limited to entity nouns, domain-specific concepts, and professional roles. This annotation strategy ensures that the dataset meets the requirements for domain knowledge extraction and model training.

After obtaining a fully machine-generated

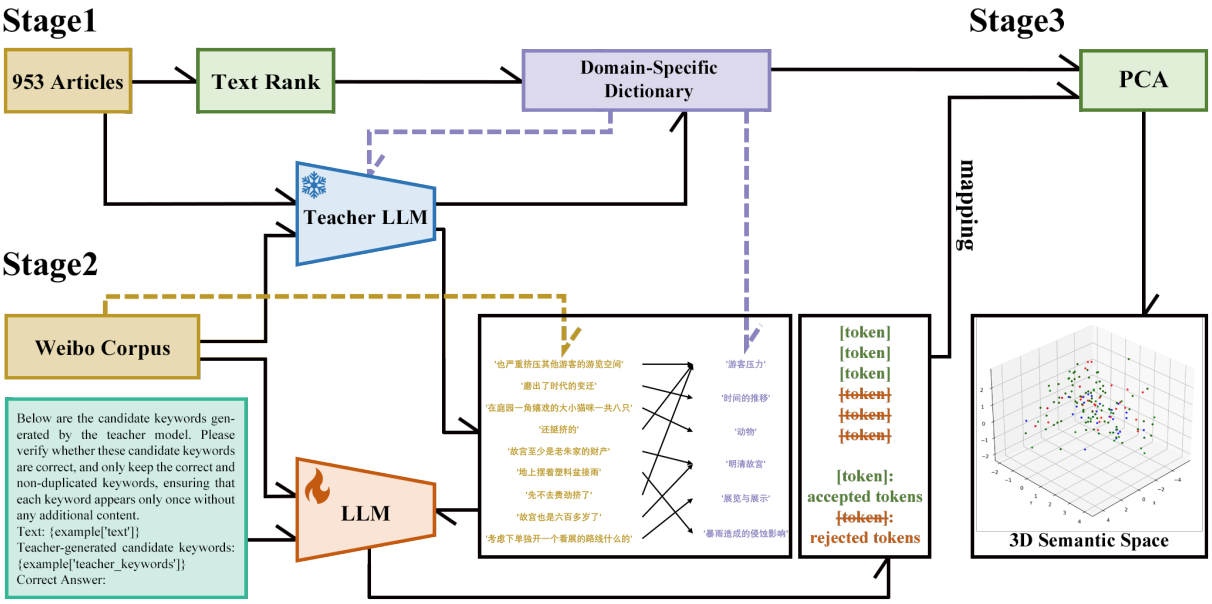

Figure 2: Overall architecture of the Generalizable Digital Cognition Framework

domain-specific dictionary, we further refined it under expert supervision, removing non-informative words (e.g., "this is," "this one," etc.) to maintain high-level domain specificity.

4 Method

4.1 Network Architecture Diagram

GDCF aims to transform everyday language into domain-specific vocabulary so that professionals in the field can more efficiently utilize everyday-language corpora. As shown in Figure 2, GDCF is composed of three modules: a domain-specific dictionary generation module, a large-model fine-tuning module, and a 3D semantic space generation module. Because no training set is available, the dictionary generation module first combines keywords extracted by a traditional TextRank algorithm with those extracted by a large model, producing a domain-specific dictionary in an unsupervised manner. Experts then add scope-based annotations—such as dynasties, emperors, and reign titles—to ensure the dictionary covers multiple dimensions of professional terms.

Next, the large-model transfer module employs an innovative teacher–student distillation approach combined with instruction tuning to enhance the smaller-parameter LLM's ability to comprehend the deeper meanings of everyday language, enabling it to accurately map everyday language to specialized vocabulary. Finally, the generated domain-specific dictionary is used by the 3D semantic space generation module to construct a semantic space, in which the transformed domain-specific vocabulary is placed. By analyzing the positions of these terms in the vector space, domain experts can better understand the semantic distances underlying everyday language.

4.2 Domain-Specific Dictionary Generation Module

This module aims to generate high-quality keywords from textual data and evaluate the predictions by comparing them with the annotations of 90 professional sentences in dataset C, supporting subsequent downstream tasks.

HanLP is one of the most advanced Chinese keyword extraction toolkits, utilizing the TextRank algorithm. In this step, we call the HanLP RESTful API to extract keywords, setting the number of keywords per sentence to $top_k = 5$, generating Dictionary A.

To further enhance the professionalism and robustness of the extracted keywords and to ensure consistency with the output of the large model fine-tuning module, we employ the same Qwen-7B-Instruct model used in the teacher model of the subsequent module. A prompt is carefully constructed to instruct the model to return only domain-specific keywords, requiring the output to be separated strictly by English commas and ensuring that all extracted keywords appear in the original sentence. The model generates output deterministically with a temperature of 0, ensuring stable results, forming Dictionary B.

Instruction Tuning	Teacher Model + Instruction Tuning
System Message: You are a high-level keyword extraction assistant in the field of cultural heritage. **User Message:** Please only return a list of professional keywords related to cultural heritage, with keywords separated solely by English commas, and do not include any other text, punctuation, or explanations, with each keyword appearing only once. All keywords must appear directly in the following sentence. Sentence: {example['text']} Keywords:	**System Message:** You are a high-level keyword validation and extraction assistant in the field of cultural heritage. **User Message:** Task: Below are candidate keywords generated by a teacher model. Please check whether these candidate keywords are correct, and only retain the correct and non-duplicated keywords, then output the final list of keywords. Text: {example['text']} Teacher generated candidate keywords: {example['teacher_keywords']} Correct answer:

Table 1: Instruction Tuning and Teacher Model + Instruction Tuning Comparison

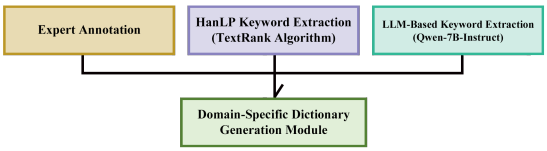

Figure 3: Domain-Specific Dictionary Module

Finally, Dictionaries A and B are merged and deduplicated, followed by expert annotation to incorporate dynasties, emperors, and reign titles, resulting in the final domain-specific dictionary.

4.3 Large Model Fine-Tuning Module

This module aims to generate pseudo-labels using a teacher model and optimize them through instruction tuning of a student model. The process consists of two steps: the teacher–student distillation process and the instruction tuning of the student model.

Step 1: We use the larger Qwen2.5-7B (Team, 2024) as the teacher model, combined with the vLLM module (Kwon et al., 2023) to generate pseudo-labels. For each input text, we construct a specific prompt, as shown in Table 1.

We then use the vLLM generation interface for sampling. The sampling parameters are set as temperature = 0.7, top_p = 1.0, and a maximum generation length of 128 tokens. The probability distribution used in this generation process follows the formula:

$$p(w_i \mid context) = \frac{\exp\left(\frac{logit(w_i)}{0.7}\right)}{\sum_j \exp\left(\frac{logit(w_j)}{0.7}\right)}$$

After processing and segmentation, the generated teacher pseudo-labels form the matrix $Q \in R^{N \times C}$.

Step 2: We fine-tune the smaller Qwen2.5-0.5B model (Team, 2024) using instruction tuning, enabling it to evaluate the teacher-generated pseudo-labels Q and retain only correct, non-redundant keywords. To achieve this, we construct an instruction-based prompt that combines the original text, the teacher-generated pseudo-labels Q, and dataset C as the final reference.

During training, we apply a label masking strategy, where the labels corresponding to the prompt tokens are set to -100, ensuring that only the target tokens contribute to the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i \in \mathcal{T}} \log P(y_i \mid x, y_{<i})$$

where $\mathcal{T}$ represents the indices of the target tokens. The student model is trained with a learning rate of 5×10^{-5}, a batch size of 1, and for 3 epochs. A custom data collector dynamically pads inputs to maintain consistent input and label lengths across batches. After fine-tuning, the student model is capable of correctly identifying the final set of accurate, non-redundant professional keywords based on both the input text and the teacher's pseudo-labels.

By integrating these two steps, the teacher model provides candidate keywords for each input text, though some noise and redundancy may be present. The student model, through instruction tuning, learns to assess and filter the teacher's output, producing a refined list of professional keywords from

everyday language. This design not only leverages the generative power of large models but also enables the small model to strike a balance between efficiency and robustness through instruction tuning, offering a novel approach to handling imbalanced and complex textual data.

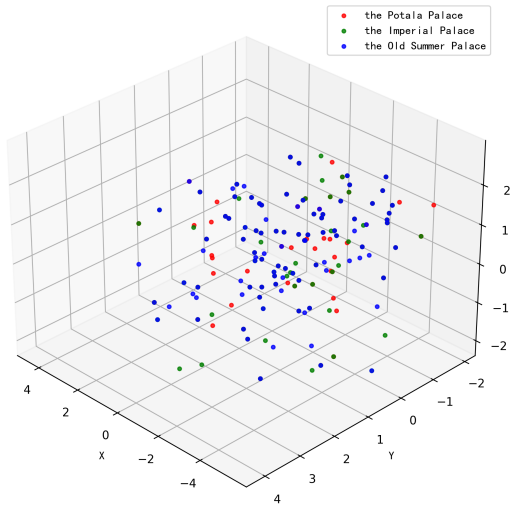

Figure 4: 3D Semantic Space

4.4 3D Semantic Space Generation Module

To illustrate the mapping relationship between everyday language and domain-specific vocabulary, we employ a keyword mapping and visualization approach using a pretrained model and dimensionality reduction techniques. First, to construct a stable vector space for keyword mapping, we obtain the domain-specific dictionary from the dictionary generation module and compute the vector representation of each keyword using the pretrained model 'shibing624/text2vec-base-chinese-paraphrase' (Xu, 2023). Since the output vectors from this model have high dimensionality, we apply Principal Component Analysis (PCA) to reduce them to three dimensions, forming a base 3D vector space.

After constructing the base vector space, we define a keyword mapping function. This function obtains the vector representation of any given keyword using the same pretrained model, then projects it into the 3D space using the trained PCA model. Formally, given a keyword w with its original vector representation E_w, its 3D mapped vector is expressed as:

$$vec(w) = PCA(E_w)$$

Next, we perform frequency statistics on sampled keywords and select the Top-N most frequent ones (e.g., $N = 100, 200, 500$). For each of these selected keywords, we apply the mapping function to convert them into 3D vectors while recording their keyword labels, frequencies, and 3D coordinates.

To visualize the spatial distribution of Top-N keywords from different CSV files, we use Matplotlib's 3D plotting tools. By setting multiple viewing angles within the same plot and using different colors to distinguish different data sources, we effectively illustrate how semantic relationships between everyday language and domain-specific vocabulary are distributed in the reduced 3D space.

5 Results and Discussion

5.1 Analysis of Domain-Specific Dictionary Generation

In the comparison of the three domain-specific dictionaries, after removing duplicate extracted terms, the statistical results show that HanLP extracted 5,024 terms, LLM extracted 3,496 terms, and the final merged and expert-processed dictionary contained 6,904 terms.

Method	Vocabulary Size
HanLP	5,024
LLM	3,496
Domain-Specific Dictionary	**6,904**

Table 2: Vocabulary size comparison of different methods. The vocabulary size represents the number of professional terms extracted by each method. The best result is shown in bold.

From these results, we observe the following: (1) Using only traditional keyword extraction methods (HanLP) can capture some domain-specific information, but due to predefined rules and algorithmic limitations, the number of extracted candidate terms is relatively small, making it difficult to comprehensively cover domain knowledge. (2) The LLM-based extraction method benefits from large-scale pretraining and demonstrates an advantage in keyword generation. However, the number of terms it generates remains relatively low, suggesting potential limitations in the model's coverage of specialized domain knowledge. (3) The merged approach with expert refinement significantly increases the size of the domain-specific dictionary, indicating that combining traditional methods with

Task	Acc	F1	Recall
Evaluation Metric	Accu.↑	F1 Score↑	Recall↑
Llama-3.1-8B-Instruct	0.140	0.135	0.149
Qwen2.5-7B-Instruct	0.482	0.410	0.374
Qwen2.5-0.5B-Instruct	0.225	0.065	0.047
Qwen2.5-0.5B-Instruct+tune	0.096	0.071	0.078
Qwen2.5-0.5B-Instruct+teacher0.5B	0.364	0.301	0.272
GDCF	0.390	0.264	0.210

Table 3: Results of LLM performance

deep learning-based generation models, along with expert validation, can leverage the strengths of each approach to produce a more comprehensive and accurate dictionary. These findings validate the effectiveness of the hybrid method in improving both the coverage and quality of domain-specific dictionaries.

5.2 Ablation Study on the Large Model Fine-Tuning Module

We conduct ablation experiments using Llama-3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Team, 2024), and Qwen2.5-0.5B-Instruct (Team, 2024), where Qwen2.5-7B-Instruct and Qwen2.5-0.5B-Instruct serve as teacher models, and Qwen2.5-0.5B-Instruct is used as the student model. The 100 annotated samples from dataset C are split into a 70:30 ratio for training and testing, and experiments are conducted in a few-shot setting. The results are shown in Table 3.

(1) Since the dataset consists of Chinese text, models are tested for their Chinese language comprehension ability. Llama-3.1-8B-Instruct performs relatively weaker, likely due to its inferior Chinese understanding.

(2) Qwen2.5-7B-Instruct, when used as the teacher LLM, achieves the best keyword generation performance, with an F1 score of 0.410. This indicates that LLMs, with their large-scale parameters and extensive knowledge, have a natural advantage in understanding text and extracting domain-specific information related to cultural heritage.

(3) When Qwen2.5-0.5B-Instruct is fine-tuned without a teacher model, accuracy (ACC) drops significantly to 0.096. This decline is mainly due to the small parameter size (0.5B), where fine-tuning in the Instruct setting makes it difficult for the model to properly understand and generate accurate outputs.

(4) The introduction of a teacher model, whether 7B or 0.5B, significantly improves the performance of instruction fine-tuning. When using 7B as the teacher model, the student model does not surpass the teacher model but achieves performance close to it, despite being constrained by the 0.5B parameter bottleneck.

(5) When using 0.5B as the teacher model, it avoids the performance degradation caused by relying solely on instruction fine-tuning while also benefiting from training two models of the same size together, leading to better results than the original model.

This ablation study strongly validates that incorporating a teacher model enhances the model's ability to understand instruction semantics during fine-tuning and mitigates interpretation biases that arise from small parameter sizes.

5.3 Cultural Cognition Interpretability Analysis in Semantic Space

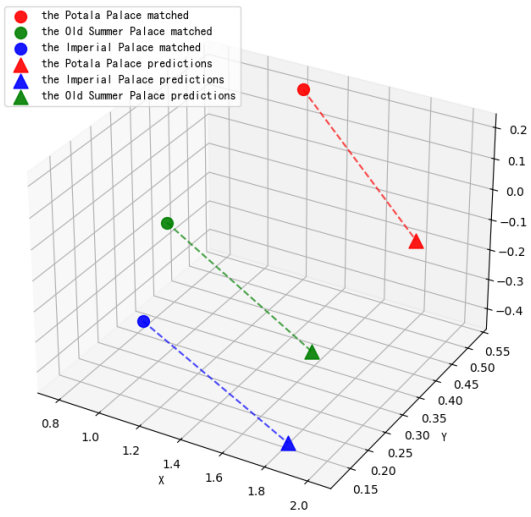

Figure 5: 3D Vector Space Distribution of Keywords from Different Cultural Heritage Categories

Figure 5 illustrates the center point distributions

of keywords from three categories (Potala Palace, Old Summer Palace, and the Imperial Palace) in a 3D vector space. The circular markers (matched) represent the center points of professional terms obtained through semantic similarity transformation from keywords extracted from everyday language using traditional methods. The triangular markers (predictions) indicate the center points of professional terms fully generated by GDCF. Dashed lines connect the two center points of the same category. Table 4 presents the Euclidean distances between these two sets of center points for each category. It can be observed that the predictions group center points are farther from the origin, suggesting that the GDCF-generated dictionary is more diverse and semantically richer.

Distance	Value
Potala Palace - Imperial Palace	0.5075
Potala Palace - Old Summer Palace	0.3645
Imperial Palace - Old Summer Palace	0.2077

Table 4: Distance Between Cultural Heritage Sites

From the Euclidean distances in Table 4, we see that the variation between center points differs across categories. For example, if the Potala Palace category shows a larger distance between matched and predictions, this may indicate that the semantic distribution of predicted keywords significantly deviates from the originally matched keyword distribution. In contrast, if the Old Summer Palace and Imperial Palace categories show smaller differences, it suggests that the predicted results for these two categories are more aligned with their original matches in semantic space.

This spatial distribution characteristic reveals a dual cultural cognition mechanism: (1) The historical narrative coupling effect causes the shared Qing Dynasty imperial architectural symbols to exert a strong semantic gravitational pull in the word vector space; (2) The political-religious hybrid function of the Potala Palace leads to cross-cultural semantic shifts of Tibetan architectural terms within Weibo corpora, which predominantly operate in a Chinese-language context.

6 Conclusion

Using UNESCO cultural heritage experts and their reports as data sources, GDCF extracts semantic features of words within the domain context, constructing a semantic space that accurately represents cultural heritage terminology.

Traditional dictionaries are compiled by linguists who define word meanings through sentence-based explanations, clarifying their actual meaning and position within the language system. However, this approach may be influenced by expert subjectivity, potentially leading to interpretations that deviate from real-world usage. In contrast, GDCF leverages fine-tuned large models to transform and define semantic attributes in the most objective manner. For instance, terms like "Acropolis of Athens" or "Baroque abbatial churches" are represented as vectors in the semantic space, directly extracted from original texts authored by cultural heritage experts. This method preserves word usage, definitions, and contextual semantics, ensuring faithfulness to expert discourse.

Moreover, since our model determines the spatial position of words and phrases through dimensionality reduction, we can identify the closest translation equivalents in a cross-lingual setting. This effectively addresses the longstanding challenge of accurately translating domain-specific dictionaries across languages.

More importantly, the model is not limited to the cultural heritage domain but exhibits broad cross-domain applicability. By integrating an advanced teacher-student model with instruction tuning, it facilitates the conversion of everyday language into professional vocabulary across different fields. This provides an efficient solution for rapidly educating both AI models and humans in specialized linguistic paradigms across various disciplines.

Limitations

The effectiveness of GDCF relies on UNESCO reports and domain expert texts, which may not fully cover cultural heritage knowledge specific to certain regions or communities. In areas with limited expert literature, the model's performance may be constrained. The teacher-student model with instruction tuning reduces dependence on large-scale annotated data, but it still requires a certain amount of high-quality labeled data, which remains a challenge in extremely low-resource scenarios.

References

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 1522–1532, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *Preprint*, arXiv:1607.01759.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko, and Thomas Hofmann. 2018. Learning and evaluating sparse interpretable sentence embeddings. *Preprint*, arXiv:1809.08621.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Ming Xu. 2023. text2vec: A tool for text to vector.

Hongyin Zhu, Hao Peng, Zhiheng Lyu, Lei Hou, Juanzi Li, and Jinghui Xiao. 2024. Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. *Preprint*, arXiv:2109.01048.

A 3D Semantic Space Visualization Results

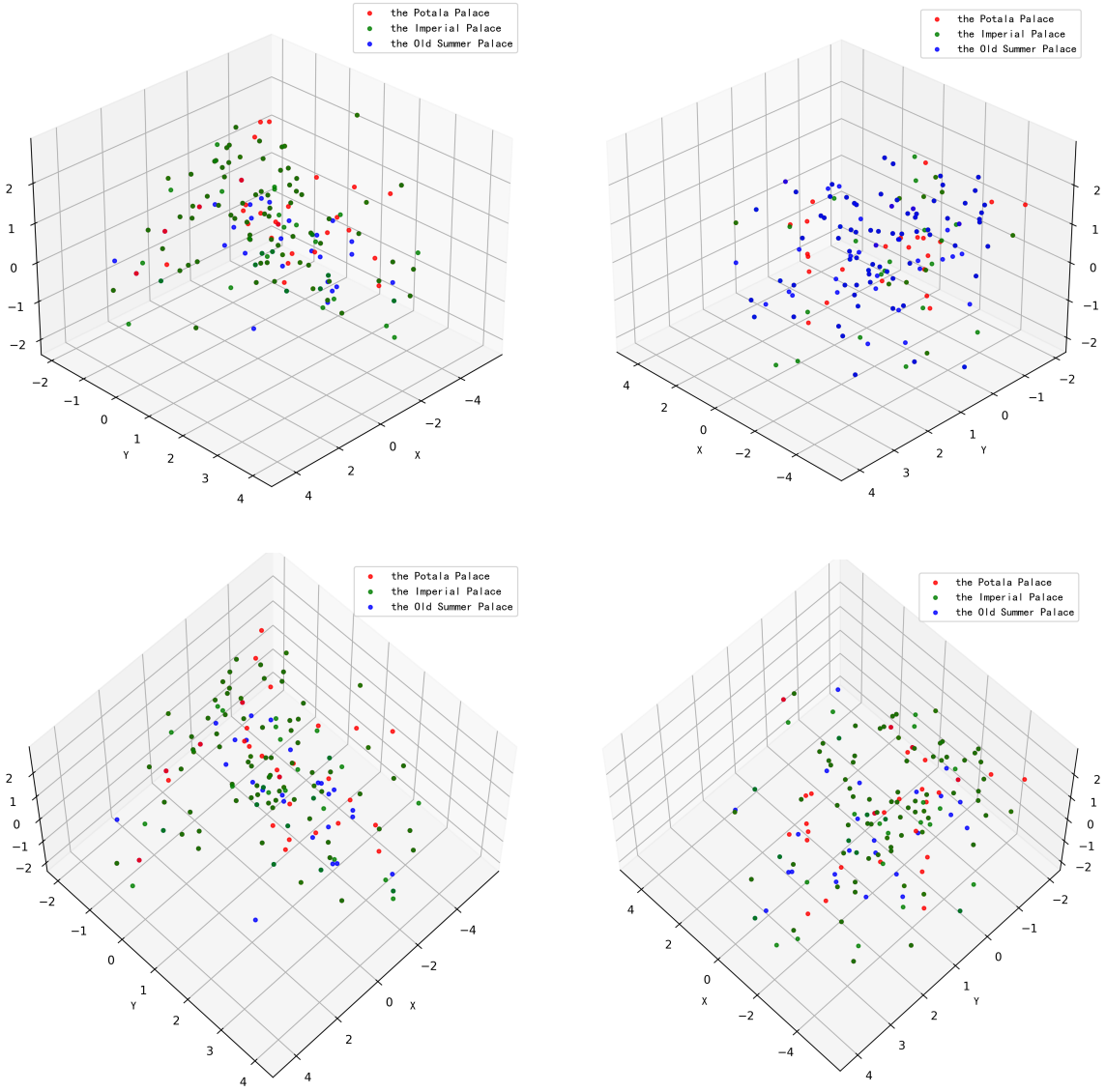

Figure 6: Top 100 Keywords in 3D Semantic Space from Different Angles

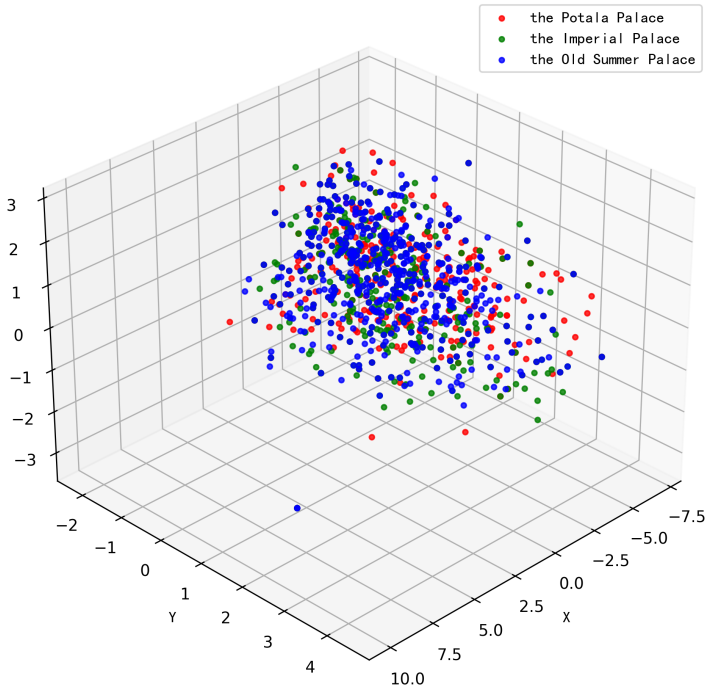

Figure 7: Top 500 Generated Keywords in 3D Semantic Space

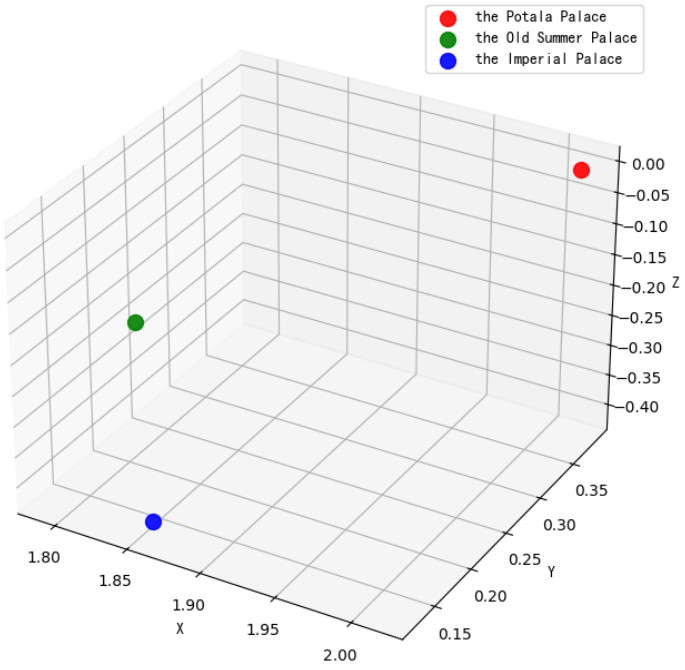

Figure 8: Comparative Analysis of Center Point Distributions for Potala Palace, Old Summer Palace, and the Imperial Palace in 3D Semantic Space