Exploring Dynamic Few-Shot Prompting for Word Sense Disambiguation in Historical Chinese

Micah Kitsunai[1] Deborah Watty[1] Shu-Kai Hsieh[1]
[1]Graduate Institute of Linguistics, National Taiwan University, Taipei, Taiwan
{r11142010, r11142012, shukaihsieh}@ntu.edu.tw

Abstract

This paper proposes a method for word sense disambiguation in historical Chinese texts using general-purpose LLMs (GPT-4o and GPT-4o-mini). The results show that the larger model performs better and few-shot examples improve performance, though the effectiveness of dynamic example selection remains unclear. The best-performing setup is applied to visualize the change in meaning of a character over approximately 3,000 years of Chinese text data, demonstrating the potential of this approach for tracking semantic evolution.

1 Introduction

The Chinese language features one of the oldest writing systems in the world, with characters that have been used for thousands of years. Despite this continuity, the meaning and usage of individual characters have evolved over time, creating a field of research focused on semantic change. Large language models have recently emerged as powerful tools for analyzing historical Chinese texts. While specifically tailored models perform well in this area, they are less flexible than general-purpose models. This paper investigates the feasibility of using GPT-4o and GPT-4o-mini, two general-purpose LLMs not specifically fine-tuned for historical Chinese, for word sense disambiguation in historical texts. To achieve this, we compare dynamic few-shot prompting, a technique designed to enhance model performance by selecting task-specific examples based on their relevance to the input query, against zero-shot and fixed few-shot approaches. We then use the best-performing setup to generate an example visualization of the change in sense frequency for a character in a corpus spanning approximately 3,000 years of Chinese text data.

2 Related Work

Dynamic few-shot prompting is a variant of few-shot prompting in which examples are selected from a database of annotated examples based on their similarity to the input query. The goal is to increase the relevance of the selected examples for the given task and thereby improve performance. Initially proposed by [1], this method has been successfully applied to various tasks, including coding [2], machine translation [3] and multimodal sentiment analysis [4].

Research on historical Chinese texts using this method remains limited. One study incorporated dynamic one-shot prompting for lexical semantic change detection [5]. Other approaches to analyzing historical Chinese with LLMs have taken different directions. For example, [6] used dynamic prompting in a translation task, relying on a Retrieval-Augmented Generation (RAG) pipeline to retrieve relevant contextual information for inclusion in the prompt.

Broader assessments of LLM performance on historical Chinese have revealed notable challenges. For example, a benchmark proposed by [7] demonstrated that even advanced models, such as ChatGLM and ChatGPT, struggle significantly more with historical data compared to modern Chinese. To address these challenges, tailored approaches have been proposed, such as GujiBERT and GujiGPT [8]. A more recent example is [9], who developed a diachronic language model for classical Chinese that achieved strong results in word sense disambiguation tasks.

3 Proposed Method

3.1 Workflow

The proposed method operates through the following steps to perform sense labeling given a character in context:

1. **Retrieval of Sense Data:** Retrieve a list of possible senses for the target character from MoeDict[1)], a Traditional Chinese dictionary.
2. **Select Few-Shot Examples**: Embed the context with `text-embedding-3-small` and select the three most similar examples from the vectorstore (see Section 3.2)
3. **Sense Labeling:** For each retrieved context, generate a prompt asking the LLM to choose a sense label for the character in context. The prompt contains:
   - The target character
   - The given context
   - The list of possible senses
   - The dynamically selected few-shot examples

   The prompt template is included in the Appendix.

3.2 Vectorstore for Dynamic Few-Shot Examples

The vectorstore contains 2300 randomly selected quotes from MoeDict, embedded using `text-embedding-3-small`, along with metadata about their origin, as shown in Table 1.

4 Evaluation

To test the efficacy of our proposed method, we compare the accuracy achieved with dynamic few-shot prompting against two simpler prompting srategies:

- **Zero-shot:** No example sentences are provided in the prompt. The model performs sense selection using only its pre-trained knowledge.
- **Fixed Few-Shot:** A fixed set of example sentences is provided in the prompt.

Like our vectorstore, our test set consists of randomly selected quotes from MoeDict. We randomly selected 312 examples to ensure a diverse representation of characters and senses.

5 Results

Figure 1 shows the accuracy for different prompt types and LLMs.

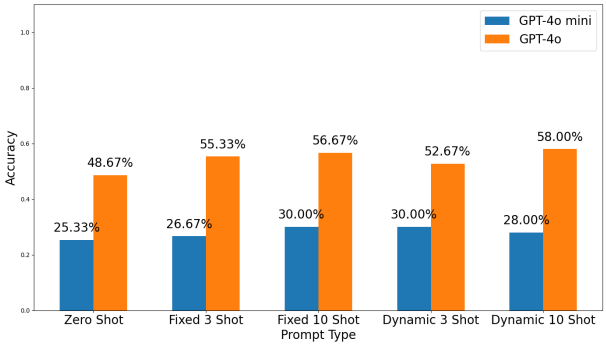

Figure 1 Accuracy Comparison by Model and Prompt Type

GPT-4o consistently outperformed GPT-40-mini across all setups. Few-shot prompting proved to be more effective than zero-shot prompting, regardless of the model used. However, increasing the number of examples in few-shot setups did not lead to significant improvements, indicating diminishing returns beyond a certain threshold. Dynamically selecting examples based on input similarity did not provide the expected performance boost. Despite this, the best overall performance was achieved with GPT-4o using dynamic 10-shot prompting, making it the most effective configuration tested.

5.1 Application to Semantic Change Visualization

A potential application of historical Chinese word sense disambiguation is visualizing how the meaning of a character changes over time. To demonstrate this, we apply the best-performing method to all occurrences of "家" (*jia*, = "home", "family", ...) in a historical Chinese corpus where each example is annotated with its corresponding dynasty [10]. We followed these steps:

1. **Retrieve Contexts:**
   - Search for target character in the Chinese historical corpus.
   - Extract the context surrounding the target characters (10 characters before and after). In this work, we experimentally retrieved 100 random examples from each dynasty in the corpus.

1) https://www.moedict.tw/

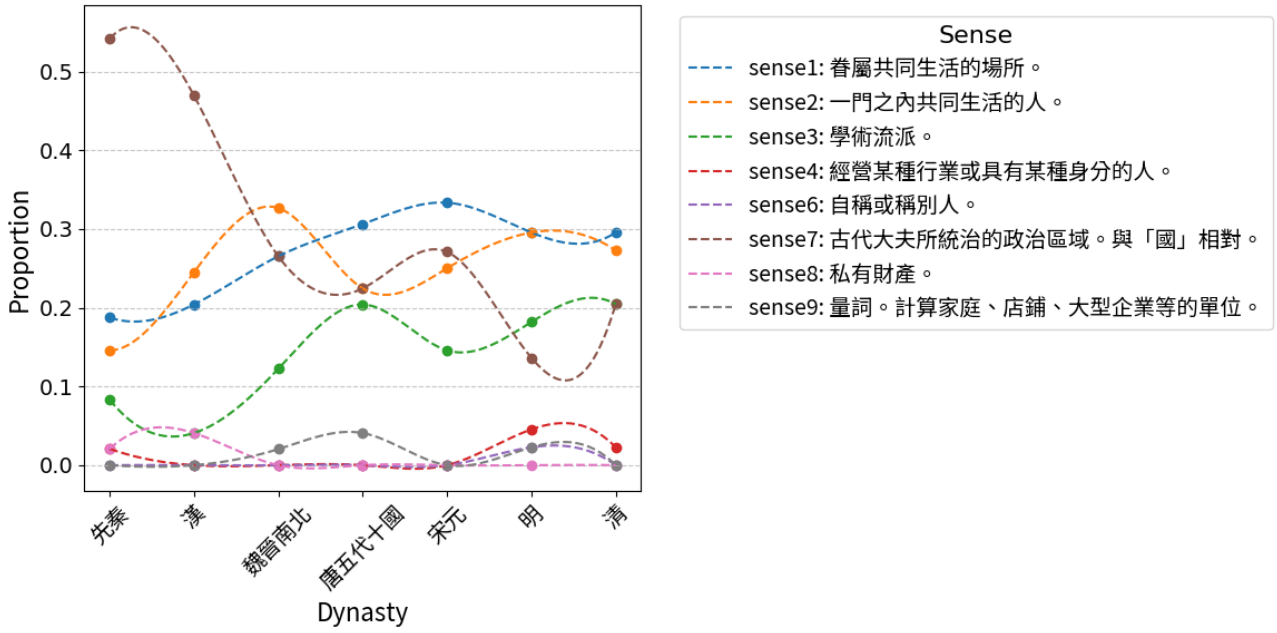

Figure 2 Distribution of the different senses of "家" over time.

- Record all <context, origin> tuples (here, origin refers to the information about the source of the given context phrase, including author, publication and dynasty).

2. **Sense Labeling:** Perform the steps outlined in Section 3.1 for each retrieved context.
3. **Visualization:**
   - For each time period in the corpus, the occurrences of each sense label for the given character are counted.
   - A plot is then generated to show the proportion of each sense label across dynasties.

The resulting semantic change plot (see Figure 2) illustrates which meanings were dominant in different dynasties, as well as how meanings emerged or disappeared over time.

6 Discussion

The overall low accuracy across configurations can be attributed to the inherent complexity of the dataset. For many characters, the set of possible senses is both large and nuanced, with semantically similar meanings often overlapping. This makes accurate disambiguation particularly difficult.

GPT-4o substantially outperforming GPT-4o-mini underscores the advantage of scale in language models. As GPT-4o is a larger model, it likely has a better capacity for encoding and differentiating semantic nuances.

This result suggests that further improvements might be achievable with even larger models in the future, pointing to an exciting direction for future research.

Dynamic few-shot prompting did not yield the anticipated improvements, which may stem from limitations in the datastore used for example selection. With only approximately 300 examples available, it is possible that less relevant examples were chosen for certain queries, reducing the effectiveness of the approach.

6.1 Limitations

This study faces two key limitations. First, like most semantic disambiguation tasks on historical Chinese, our method relies on predefined senses for characters. We use the MoeDict definitions, which, while comprehensive, may not fully capture all meanings a character could have held in the past. This may limit the model's ability to disambiguate less common or historically nuanced senses. Second, our datastore is relatively small, which likely impacts the relevance of the few-shot examples selected during prompting. A larger datastore might provide examples that are better aligned with the input queries, potentially improving performance.

Finally, in addition to the limitations of our method, readers should note that the setup used to create the semantic change plot in Figure 2 achieved only 58% accuracy in the experiment. The plot is included solely to illustrate a po-

tential application, assuming necessary improvements to the methods are made.

6.2 Future Work

Future efforts will focus on expanding the datastore to include a broader range of examples, which could further improve the relevance of few-shot prompting. Additionally, we aim to develop a tool that allows users to input any character and generate a visualization of its senses over time, similar to the one shown in Figure 2.

7 Conclusion

This study explored the potential of dynamic few-shot prompting with general-purpose LLMs (GPT-4o and GPT-4o-mini) for word sense disambiguation in historical Chinese. The results indicate that model size plays a crucial role, with GPT-4o significantly outperforming GPT-4o-mini. However, the method has notable limitations. Overall accuracy was low, with the best-performing setup—GPT-4o with dynamic 10-shot prompting—achieving only 58%. Using fewer examples or skipping dynamic selection did not result in drastically worse performance, leaving the advantages of dynamic prompting unproven. This may stem from the small datastore, which could limit the availability of relevant examples for some inputs.

Overall, this study highlights both the promise and the limitations of dynamic prompting with LLMs for this task. Future work should focus on expanding the datastore to fully realize the potential of this approach.

References

[1] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3?, 2021.

[2] Dustin Dannenhauer, Zohreh Dannenhauer, Despina Christou, and Kostas Hatalis. A case-based reasoning approach to dynamic few-shot prompting for code generation. In **ICML 2024 Workshop on LLMs and Cognition**, 2024.

[3] Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. Adaptive machine translation with large language models. **arXiv preprint arXiv:2301.13294**, 2023.

[4] Li Yang, Zengzhi Wang, Ziyan Li, Jin-Cheon Na, and Jianfei Yu. An empirical study of multimodal entity-based sentiment analysis with ChatGPT: Improving in-context learning via entity-aware contrastive learning. **Information Processing & Management**, Vol. 61, No. 4, p. 103724, 2024.

[5] Zhengfei Ren, Annalina Caputo, and Gareth Jones. A few-shot learning approach for lexical semantic change detection using GPT-4. In **Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change**, pp. 187–192, 2024.

[6] Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. Large language models for classical Chinese poetry translation: Benchmarking, evaluating, and improving, 2024.

[7] Yixuan Zhang and Haonan Li. Can large language model comprehend ancient Chinese? a preliminary test on aclue. **arXiv preprint arXiv:2310.09550**, 2023.

[8] Dongbo Wang, Chang Liu, Zhixiao Zhao, Si Shen, Liu Liu, Bin Li, Haotian Hu, Mengcheng Wu, Litao Lin, Xue Zhao, et al. GujiBERT and GujiGPT: Construction of intelligent information processing foundation language models for ancient texts. **arXiv preprint arXiv:2307.05354**, 2023.

[9] Yuting Wei, Meiling Li, Yangfu Zhu, Yuanxing Xu, Yuqing Li, and Bin Wu. A diachronic language model for long-time span classical Chinese. **Information Processing Management**, Vol. 62, No. 1, p. 103925, 2025.

[10] Micah Kitsunai, Deborah Watty, and Shu-Kai Hsieh. Building a semantic search platform for exploring historical Chinese corpora. じんもんこん 2024 論文集, Vol. 2024, pp. 241–246, 2024.

Character	Context	Origin	Sense	Possible Senses	Embedding
家	少小離家老大回，音無改鬢毛衰。」	唐：賀知章．回偶書詩二首之一	眷屬共同生活的場所	[居住。, 眷屬共同生活的場所, 家中的。]	[0.345, -1.4235, 0.2345.....]
...	...	...	...	...	

Table 1 Example entries from the vectorstore.

A Appendix

Prompt Template

You are an expert in ancient and modern Chinese linguistics. Given a Chinese character, its context, and possible sense labels, your task is to identify the sense label that best fits the character's usage in the given context. Use the examples provided to guide your decision-making process.

Examples:
Character: {example_char_1}
Context: {example_context_1}
Origin: {example_origin_1}
Possible Sense Labels: {example_sense_labels_1}
Correct Sense Label: {correct_sense_1}

Character: {example_char_2}
Context: {example_context_2}
Origin: {example_origin_2}
Possible Sense Labels: {example_sense_labels_2}
Correct Sense Label: {correct_sense_2}

Character: {example_char_3}
Context: {example_context_3}
Origin: {example_origin_3}
Possible Sense Labels: {example_sense_labels_3}
Correct Sense Label: {correct_sense_3}

Question:
Character: {character}
Context: {context}
Origin: {origin}
Possible Sense Labels: {sense_labels}

Which of the sense label best fits this usage of the character? Respond with the single most appropriate sense label in the following format:

```
{
 "label": string // most appropriate sense label for {character} in {context}
}
```