Parsing Historical Job Titles via LLMs to Analyze Social Mobility in Late Imperial China

Yue Yu*[1], Jun Chen[2], Michael Yan Hon Chung[1], and Cameron Campbell[1]

[1]Hong Kong University of Science and Technology
[2]Renmin University of China

Abstract

The digitization of Chinese historical career records has provided millions of data points for the study of social mobility. However, the unstructured nature of Chinese bureaucratic job titles—where an official's functional office is conflated with honorary ranks, duty assignments, and prestige titles—has created a substantial barrier to accurate quantitative measurement, which requires a clear distinction between substantive power and symbolic status. We address this bottleneck by fine-tuning a Small Language Model (Qwen3-8B) on thousands of annotated pairs of job titles in the Qing dynasty to decompose these strings into a structured historical schema. We demonstrate that this specialist model outperforms commercial Large Language Models (LLMs) in both accuracy and cost-efficiency. By applying this parser to the full China Government Employee Database-Qing *Jinshenlu* (CGED-Q JSL) with over four million records, our empirical analysis reveals a sharp "inflation of honors" following the Taiping Rebellion (post-1864), quantitatively identifying the demographic sub-groups that leveraged financial capital to acquire symbolic status as the dynasty, desperate for revenue, resorted to the sale of honors. Our approach can be applied to other sources, including from historical Korea, Japan, and Vietnam, where job titles or other information about status is written in Chinese characters.

1 Introduction

The abundance of career records from Qing China offers a unique opportunity to study long-term social mobility and stratification. The China Government Employee Database-Qing *Jinshenlu* (CGED-Q JSL) alone contains over four million records spanning from 1760 to 1912 (Chen et al., 2020). However, quantitative analysis is currently constrained by the unstructured nature of the critical variable defining an official's career: the job title.

As illustrated in Figure 1, a single entry often contains a complex mixture of functional offices, prestige titles, and duty assignments (Hucker, 1988). To a computer, this is merely a string of characters; to a historian, it represents a precise hierarchy of power. For example, "Vice Minister" (*Shilang*) is typically a substantive high office. However, when the string reads "Governor of Guizhou, Vice Minister," the same term functions purely as an honorary concurrent title used to elevate the Governor's rank.

*yue.yu@connect.ust.hk

Figure 1: Decomposition of a complex Qing Dynasty job title string. The raw text must be disentangled into distinct administrative concepts—Symbolic Honors, Functional Power, and Jurisdiction—to be analytically useful.

Traditional rule-based parsers based on keyword matching or regular expressions (such as MARKUS by De Weerdt (2020)) struggle to distinguish these context-dependent nuances, particularly given the "long tail" of transcription errors and irregular abbreviations found in the historical record (Campbell and Chen, 2022). The advent of Large Language Models (LLMs) offers a solution, as these models possess the semantic understanding of natural language (Bail, 2024). However, relying on commercial APIs presents challenges regarding reproducibility, data privacy, and prohibitive costs when processing millions of records (Lin and Zhang, 2025).

In this study, we introduce a domain-specific solution: a fine-tuned local Small Language Model (SLM). We demonstrate that by teaching a smaller model the specific "grammar" of Qing bureaucracy via Supervised Fine-Tuning (SFT), we can achieve high-accuracy decomposing a long job title to different basic elements.

2 The LLM Parser

To address the limitations of both rigid rules and expensive generalist models, we developed a reproducible workflow to fine-tune a local LLM.

2.1 Schema Design

We defined a rigorous JSON schema that maps unstructured text to sociological variables. Unlike simple entity extraction, this schema forces the model to categorize information based on administrative function, decomposing the raw string into five key dimensions with 15 fields in total:

- **Substantive Power:** Identifies the functional *Core Office* that determines the official's current daily authority. For officials on temporary commissions, it also extracts their *Base Office* (e.g., distinguishing a "Ministry Secretary" serving as a "Superintendent").
- **Duty Assignment:** Isolates temporary *Assignments* from stable bureaucratic posts.
- **Symbolic Capital:** Isolates prestige markers including *Honors* (e.g., Peacock Feathers), *Noble Titles*, *Ceremonial Titles*, *Warrior Titles*, and *Additional Titles* (Brevet Ranks).
- **Tenure & Qualification:** Captures the official's *Administrative Status*, distinguishing between substantive appointments and Acting statuses, as well as *Qualification Titles* (such as "Promised Promotion").

- **Department, Jurisdiction, & Garrison:** Parses the hierarchical location of the post, separating *Central*, *Local*, and *Military* departments from specific *Jurisdictions* and *Garrisons*.

2.2 Data Annotation and Augmentation

A major challenge in training AI for history is the scarcity of labeled data. We addressed this by constructing a training dataset of 3,000 pairs using a hybrid approach:

1. **Expert Annotation (Real Data):** We manually annotated 2,000 records randomly sampled from the CGED-Q JSL as the "ground truth" for standard bureaucratic patterns.
2. **Synthetic Augmentation (Simulated Data):** To make the model robust against rare cases, we programmatically generated 1,000 synthetic examples. We applied specific mutation strategies, such as removing geographic context and recombining attributes, to force the model to learn the semantic logic of the titles rather than merely memorizing common strings.

2.3 Model Fine-Tuning

We selected the **Qwen3-8B** as the base model (Yang et al., 2025). We utilized **LoRA** (Low-Rank Adaptation) for parameter-efficient fine-tuning (Hu et al., 2022). This approach allowed us to update only a small fraction of the model's weights specifically for the parsing task, reducing computational requirements. The model was trained on the 3,000-pair dataset using an instruction-tuning format.

3 Model Performance

We evaluated the model on a hold-out test set of 200 distinct records. We compared our Fine-Tuned Qwen3-8B against a leading commercial generalist model, DeepSeek V3.2 (DeepSeek-AI, 2024), with exactly the same prompt.

Table 1: Performance Comparison: Extraction Accuracy (Exact Match/F1)

Target Field	**Fine-Tuned Qwen3-8B**	**DeepSeek V3.2 (Zero-Shot)** [a]
Core Office (Power)	**90.00%**	**68.50%**
Base Office	93.00%	87.50%
Office Status	95.50%	63.50%
Symbolic Capital (Avg.)[b]	98.22%	93.10%
Assignments	76.13%	64.25%
Geography & Depts (Avg.)[c]	96.76%	77.15%
Overall Schema Average	**94.64%**	**80.88%**

[a] *Evaluated in February 2026 using DeepSeek API.* [b] *Average of Honors, Noble, Ceremonial, Warrior, and Additional Titles.* [c] *Average of Local, Central, and Military Departments. Overall Average is calculated across all 15 fields.*

The results highlight the necessity of domain-specific fine-tuning. While the generalist DeepSeek model achieved a respectable overall average of 80.88%, it struggled significantly with

the most historically critical variable, *Core Office* (68.50%). This suggests that generalist models often confuse substantive offices with other titles when the syntax is complex. In contrast, our fine-tuned specialist model achieved **90.00%** accuracy on *Core Office* extraction, validating it as a reliable instrument for measuring social mobility.

4 Future Works

With the parsing problem resolved, our immediate next step is to deploy this model across the full 4.4 million records of the CGED-Q JSL. This will enable the first systematic quantitative analysis of the "Inflation of Honors" in the late Qing, the results of which we will present at the SSHA meeting in Atlanta. In the last half of the 19th century, the Qing state became increasingly desperate for revenue, and began selling honorary titles that for government officials were listed alongside their actual job titles in official documents.

Preliminary results suggest a structural change beginning in the 1870s, where the number of individuals holding high-ranking titles (via *Additional Titles* and *Honors*) surged. By leveraging the structured data on titles, including purchased honorary titles, we aim to measure the rate of this inflation and map the geographic and social origins of the individuals who purchased honors. This will help identify the social groups that thought it necessary or at least worthwhile to purchase such honors as a means of showcasing their status.

Moreover, we will open-source the code and the fine-tuned model to enable everyone to efficiently parse Chinese historical bureaucratic job titles for their own research.

References

Bail, C. A. (2024). Can generative ai improve social science? *Proceedings of the National Academy of Sciences 121*(21), e2314021121.

Campbell, C. and B. Chen (2022, Sep.). Nominative linkage of records of officials in the china government employee dataset-qing (cged-q). *Historical Life Course Studies 12*, 233–259.

Chen, B., C. Campbell, Y. Ren, and J. Lee (2020). Big data for the study of qing officialdom: The china government employee database-qing (cged-q). *Journal of Chinese History 4*(2), 431–460.

De Weerdt, H. (2020). Creating, linking, and analyzing chinese and korean datasets: Digital text annotation in markus and comparativus. *Journal of Chinese History 4*(1), 17–42.

DeepSeek-AI (2024). Deepseek-v3 technical report.

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hucker, C. (1988). *A Dictionary of Official Titles in Imperial China*. Southern Materials Center.

Lin, H. and Y. Zhang (2025). Navigating the risks of using large language models for text annotation in social science research. *Social Science Computer Review 0*(0), 08944393251366243.

Yang, A., A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu (2025). Qwen3 technical report.