Can Large Language Models Understand Chinese Genealogical Images?

12.2025 Chiun-Hau You

Introduction

Chinese genealogical records (家谱) preserve centuries of family history through a combination of text and visual relationship diagrams, yet their complex layouts make them difficult to digitize with existing OCR systems. While recent advances in multi-modal LLMs show potential, there is no study applying them in the study of Chinese genealogy. Our goal is to test the ability of state-of-the-art LLMs in understanding and interpreting traditional Chinese genealogical images and explore how they can be used in digital humanities.
We developed a program that generates synthetic genealogical images mimicking authentic Chinese genealogical images, and asked LLMs to answer two types of questions: 1) predict given two figures' relationship in the family, and 2) identify all figures of a certain relationship of a given figure's name. We generated a total of 60 genealogical images of various dimensions, and tested 22 questions against each image using Gemini 2.5 Flash and Gemini 3 Flash.

Research Questions

1. How well does state-of-the-art LLM understand Chinese genealogical images?
2. What factors might affect their performance? Do they perform better on specific relationship types or tree size?

Background: Pilot Test and the Limitation of Authentic Data

During the pilot study, we obtained a few authentic images from the author's own family genealogical record and manually tested them with different LLMs and prompts. We initially thought the strongest LLMs, such as GPT-5, might be capable of reconstructing the family tree digitally, similar to what OCR systems do but with a structural tree output, but the result turned out to be rather disappointing. They consistently made critical mistakes, such as failure to recognize key figures in the family tree or misinterpreting sidenotes as persons, leading to completely useless results. Recognizing the limitation of this approach, we tried

another prompting approach by asking LLMs to predict the relationship of two figures in the tree (e.g. What is the relationship between [Person A] and [Person B]?) or to find figures that satisfy a specific type of relationship (e.g. Who are all the [GRANDCHILDREN] of [Person B]?). This approach yielded relatively accurate results but the accuracy still varied significantly between different images and different relationship types.

The inconsistent results made us interested to know what factors might have affected LLM's performance in understanding Chinese genealogical records, how reliable they are in larger computational settings, and how to get the most fruitful results from them in future digital humanities study.

To answer this, we need to acquire more annotated genealogical data to test both quantitatively and qualitatively. However, due to the limited size of the authentic data and the extensive human effort needed to annotate them, we decided to generate synthetic data which replicate the style and structure of the authentic data. In this way, we can control variables for each family tree (e.g. number of generations, size of a generation, etc), generate them in batches, and have the images and annotation ready at the same time.

Data Synthesis

1. Generating family trees

We developed a Python program to generate random family trees in JSON format. It takes two parameters: number of generations (N), and average size of a generation (S). This gives us an intuitive sense over the size of a generated family tree. In total, we generated 60 synthetic family trees whose N ranges from 4 to 7 and S ranges from 3 to 7, leading to 20 different sets of combinations with each combination having 3 different outputs. To mimic the randomness and asymmetry of real family tree structure, the program is designed to ensure at least one branch that reaches N generations while another major branch ends 1 to 2 generations earlier. It is important to note that, we developed the rules based on intuition and did not evaluate how "real" those synthetic data is compared with authentic data, and we do not intend to do so. We recognized that by imperatively coding those seemingly random rules, we actually introduced significant bias to the data, since genealogy is never random. However, considering the goal of the study is to evaluate how well LLM understands this specific type of data and to discover factors affecting their performance, a controlled synthetic data can be treated as a convenient proxy, thus justified the research design. We

focus more on the practical and visual aspects of such documents, not their representativeness of genealogy.

2. Generating images

Another separate web application was developed to render the JSON files into the actual family tree images. In this way, we have the synthetic images and ground truth annotations in JSON ready.

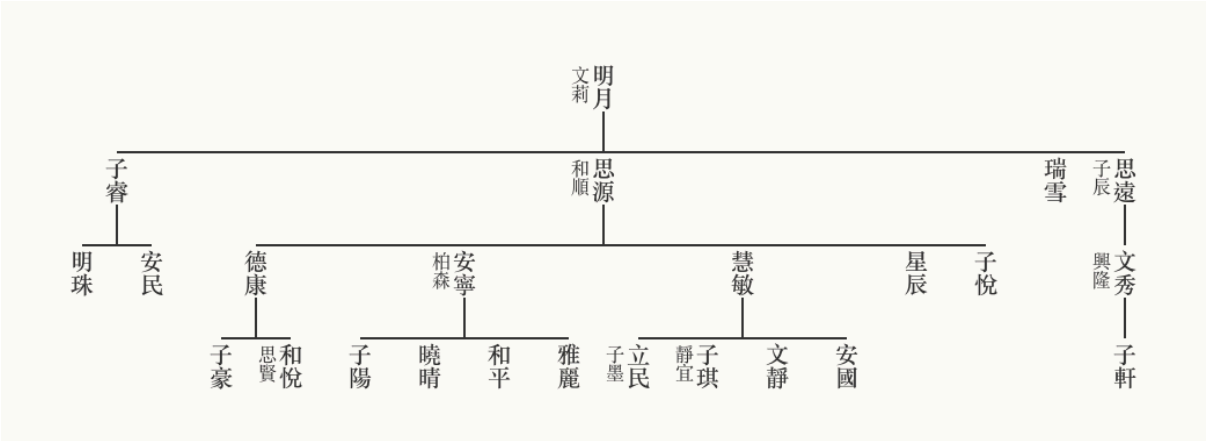


Figure 1: One of the synthetic genealogical image of N=4 and S=6 (24 nodes in total)

Task Scenarios

Next, we designed two “pathfinding” task scenarios for LLM to perform on each image. In Task 1, we asked LLM to predict the relationship between two given figures presenting in the tree. In Task 2, we asked LLM to “reverse lookup” to find all the figures of a given relationship type of a given figure. Each image was assigned 11 questions of Task 1 and another 11 questions of Task 2. All the test questions and correct answers were derived from the JSON files, ensuring the feasibility of answering the questions.

We argue that these task scenarios are perhaps most fundamental and useful for Chinese genealogy historians since Chinese kinship terms are notoriously granular (e.g. distinguishing between a mother’s brother and a father’s brother), it requires knowledge in classical Chinese language and context reasoning of the spatial hierarchy and sometimes reading the annotations to correctly infer the relationship, which is often the first thing one should determine in order to make any further interpretation. On the other hand, during the pilot test, we also observed that these “pathfinding” tasks seem to be often the first underlying steps a LLM would take in its reasoning process when dealing with a larger task involving the family tree. This observation, while not always right, made us believe that

testing LLMs' performance in "pathfinding" tasks has its guiding value for future studies involving more advanced interpretation tasks, and it also fits in the constraints of our synthetic data and limited timeframe.

Task 1: Predict the type of relationship of given two figures

In the task, a LLM is prompted with a family tree image to predict the relationship between [Person A] and [Person B], where Person A and Person B are guaranteed to be presented in the family and their relationship is one of the 11 main types. This task is evaluated as binary correct or incorrect and reported as accuracy rate in percentage. The following text snippet is the textual part of the prompt.

```
Context: This is a Chinese family tree.
Task: Please answer what is the relationship between {person_a} and {person_b}?
Note:
  - Answer in the form of "Person A is Person B's [relationship]" but return only the relationship.
  - Only reply the relationship as either CHILD, SPOUSE, PARENT, SIBLING, GRANDCHILD, GRANDPARENT, GREAT_GRANDCHILD, GREAT_GRANDPARENT, UNCLE_OR_AUNT, NEPHEW_OR_NIECE, COUSIN.
  - If you can not find the persons in the family tree, just reply "NOT_FOUND"
  - If you find the relationship but not sure, reply "OTHER"
```

Task 2: Find all figures of given type of relationship of a given figure

In task 2, we reversely asked LLM to find all the figures of specific relationship type from the point of view of Person B in the Task 1 questions set. This guarantees that at least one Person A exists in the family. Since this task is similar to a classifier, we report the result using precision, recall, and F1-score. The following text is the textual part of the prompt.

```
Context: This is a Chinese family tree.
Task: Please answer who are ALL the {relationship_type} of {person_b}?
Note:
  - Answer with a simple comma-separated list of names, e.g. "Name1, Name2".
  - If there is only one, just return the name, e.g. "Name1".
  - If you can not find any persons, reply "NOT_FOUND".
```

Results

We initially only planned to test against Gemini 2.5 Flash, since it is the only cost-effective LLM suitable for the tasks that stand out during the pilot test. However during the writing of the report, a same tier but more advanced model Gemini 3 Flash was released. We then tested the newer model with the exact same setup and got surprisingly much better results than the Gemini 2.5 Flash.

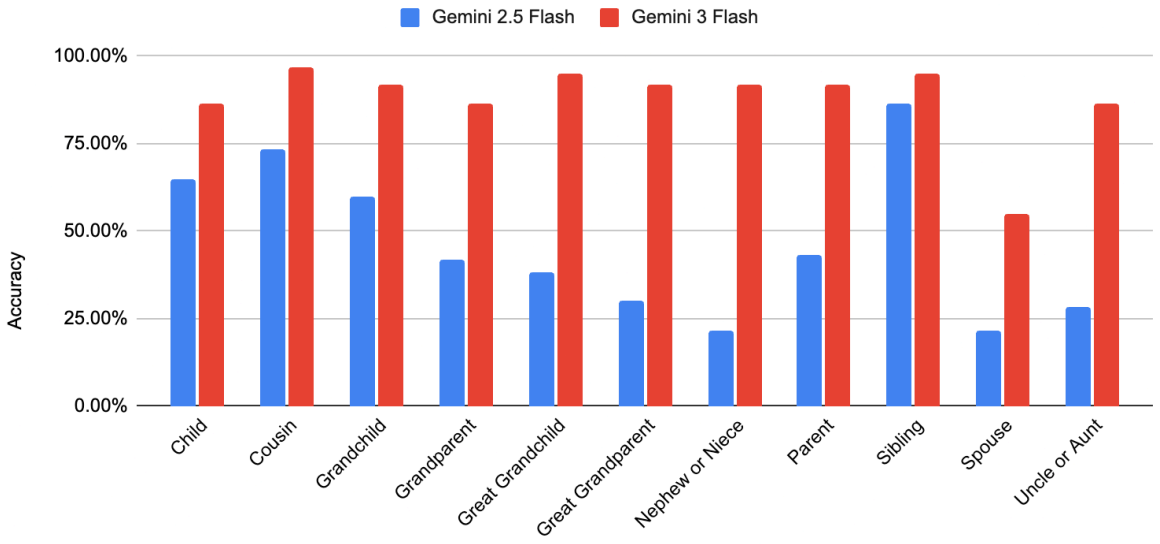
Task 1

Our result (Table 1) shows that Gemini 2.5 Flash answered almost half of the questions (46.36%) correctly, with its performance decreasing as the number of generations increases (max=54.55, min=40.61, σ=5.99%). The latest Gemini 3 Flash nearly doubled the performance, reaching an average accuracy of 88.03% and is slightly more resilient to generation depth (σ=4.63%).

Table 1: LLMs' performance in task 1 on average and grouped by generation depth (G)

	Accuracy (%)				
	Average	G=4	G=5	G=6	G=7
Gemini 2.5 Flash	**46.36%**	54.55%	46.67%	43.64%	40.61%
Gemini 3 Flash	**88.03%**	93.33%	86.67%	89.70%	83.42%

As for the accuracy by relationship type (Figure 2), we found that Gemini 2.5 Flash performed relatively well for siblings, cousins and grandchildren, but performed significantly worse on spouse, nephew or niece, and uncle or aunt, showing high inconsistency (max=86.67%, min=21.67%, σ=21.89%). Gemini 3 Flash, on the other hand, performed consistently well among most types, with 10 of the 11 types reaching an accuracy above 85% (max=96.67%, min=55.00%, σ=11.49%). Interestingly, for both models, spouse remains the most difficult relationship to identify correctly. This might be related to the fact that, instead of being a standalone node, spouse is often annotated in a small text on the left of a male figure in Chinese genealogy, also in our synthetic data, which poses challenges to LLMs.



Task 2

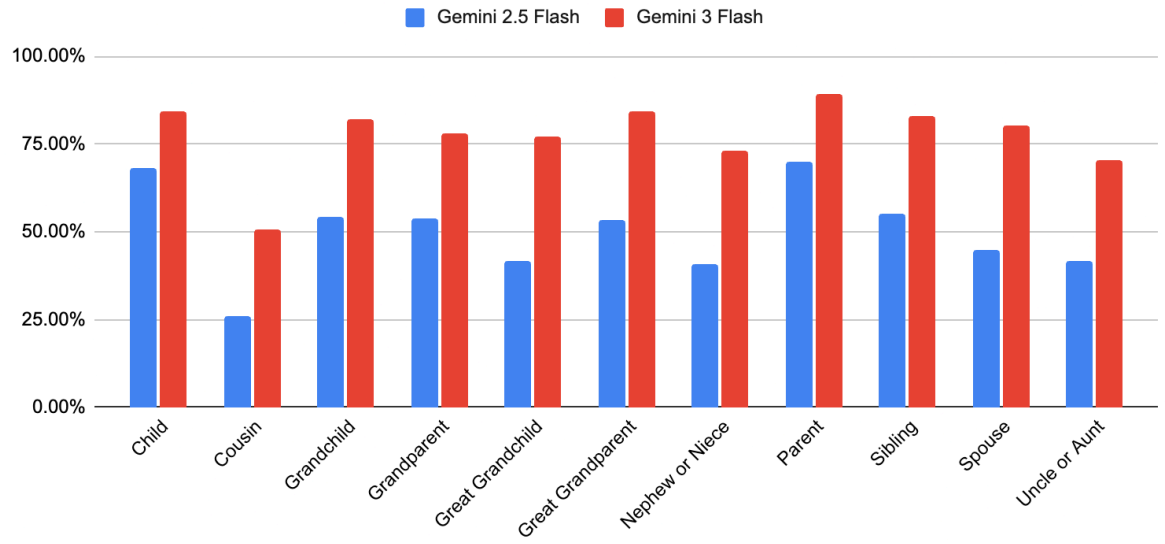
Similar to Task 1 result, Gemini 2.5 Flash performed moderately well, getting an overall F-1 score of 49.94%. However, as the family tree grows from 4 to 7 generations, its F1-score drops by a massive 22.98% (from 62.99% to 40.01%), showing the model's reasoning ability breaks down quickly as the logical distance increases. The newer Gemini 3 Flash model's F1-score (77.38%) is significantly higher and only experiences a minor 6.52% drop in F1-score across the same range, showing the new model's stability against increased complexity.

Table 2: LLMs' performance in task 2 on average and grouped by generation depth (G)

		Overall	G=4	G=5	G=6	G=7
Gemini 2.5 Flash	F1-score	**49.94%**	62.99%	50.92%	45.84%	40.01%
	Precision	47.55%	60.57%	48.31%	43.26%	38.06%
	Recall	59.43%	74.35%	61.80%	54.44%	47.12%
Gemini 3 Flash	F1-scroe	**77.38%**	80.16%	77.24%	78.49%	73.64%
	Precision	75.86%	78.28%	74.65%	76.45%	74.06%
	Recall	83.92%	86.99%	85.03%	86.26%	77.42%

As for the performance against each relationship type, we see similar patterns from both models, with Gemini 3 Flash performing relatively stable across all relationship types. Both models performed best in reasoning along direct lineages (Parents/Grandparents) but worse

in collateral lineages (Uncles/Aunts/Cousins). For example, both models struggle with identifying all cousins, uncle or uncles, with the better Gemini 3 Flash model having a F1-score on cousins for only 50.44% for example.



Discussion

Key insights

Comparing the results of the two tasks, we found that in general, LLMs' performance on all tasks would likely decrease as the complexity of the family tree grows, although the trend is less obvious for the more advanced model. This suggests that, to maximize the performance of LLMs in understanding genealogical images, keeping the size of a family in the image as small as possible (e.g. by cropping the image to keep only the relevant branches during image preprocessing) is generally preferred.

For relationship type inference, we observed that those types involving spouse lookup are more likely to cause error. As the result reveals that, among those false predictions on spouse relationship, sibling is the top false prediction. This mistake is quite understandable as in Chinese genealogy, a spouse appears as a smaller text next to the male figure, making it more difficult to distinguish between spouse and siblings as both all appear as names distributed side by side under a horizontal line, with font size and spacing being the only differences, which is easily affected by design choice. This suggests that LLMs might struggle to understand the semantics of the subtle visual differences. To mitigate this, we hypothesize that explicitly instructing the LLM to "look closely" at these subtle visual

differences or providing a context detailing those printing conventions specific to the dataset via prompt engineering might help.

For finding all targets of specific relationship types, it is clear that LLMs struggle with collateral lineages involving looking up and hoping across multiple sub trees (e.g. cousins, uncles or aunts). We recognized that these multi-choice questions might be naturally more challenging for LLMs, hence a lower F1-score is expected. However, to explain the exceptionally low score on “cousins”, we hypothesize that the error rate was amplified by what we already pointed out previously that LLM might have falsely included cousins’ spouses as cousins. This is revealed by its low precision (40.55%) and high recall (83.33%) on prediction of cousins. For other collateral lineages, like “uncles or aunts” and “nephew and niece”, the error is not that significant probably due to their gender-inclusive definition. For those two types, there is no need to exclude the spouses, hence they are not sensitive to the spouse lookup issue as experienced in the prediction of cousins.

Overall, our results show that the latest Gemini 3 Flash model is capable of performing basic pathfinding tasks against synthetic Chinese genealogical images with high accuracy, with an accuracy of 88.03% for Task 1 and a F1-score of 77.38% for Task 2. While tree complexity has a slightly negative impact on its performance, the main obstacle remains in tasks involving distinguishing subtle visual differences, such as determining spouse and siblings, who only differ in text styles. Future studies would benefit from experimenting with more fine-tuned prompts to provide LLM more context to navigate through dataset-specific design conventions.

From “pathfinding” to extracting meanings

In the study, we only focused on “pathfinding” tasks based on an assumption that LLM’s pathfinding skill is fundamental to any tasks requiring genealogical image reasoning, and a practical consideration of having limited annotated authentic genealogical images.