Unlocking Legal Insights:

Applying Large Language Models to Analyze Divorce Judgments in China

Emma Zang, Yale University
Yishi Yin, Yale University
Zitong Wang, Chinese University of Hong Kong
Han Zhang, Brown University

Abstract

Legal judgments contain rich demographic-relevant information, yet their volume, complexity, and unstructured nature have made them an underutilized resource in demographic research. This study harnesses large language models (LLMs) to systematically analyze 3,602 Chinese divorce judgments from Zhejiang Province, drawn from a larger dataset of 72,102 cases nationwide between 2009 and 2016. Unlike traditional approaches reliant on manual coding or rule-based methods, fine-tuned LLMs, such as GPT-4o-mini, demonstrate superior accuracy in extracting key demographic variables—including gender, custody claims, and property division—by capturing implicit social cues and resolving ambiguities in legal text. Comparing ten computational strategies, we find that fine-tuned decoder-only models outperform traditional NLP methods, but performance also depends on factors such as hyperparameter tuning, preprocessing techniques, and linguistic context. This study is the first to apply LLMs to Chinese divorce judgments in demographic research, providing both empirical and methodological advancements. Our results highlight the scalability and efficiency of AI-driven text analysis, offering a robust alternative to resource-intensive manual coding. By systematically comparing rule-based methods, traditional machine learning, and transformer-based models, we offer a framework for selecting optimal classification techniques based on text structure and computational trade-offs. More broadly, our findings demonstrate how LLMs can expand the analytical scope of demographic research by processing vast unstructured datasets. While this study focuses on a sample of 3,602 cases, our method could efficiently code over 1.5 million legal judgments across China, paving the way for large-scale investigations of family law, socioeconomic disparities, and policy impacts. As AI continues to evolve, its integration into demographic research presents new opportunities to extract insights from legal texts at an unprecedented scale.

Introduction

Legal frameworks play an increasingly central role in shaping demographic outcomes, influencing patterns of fertility, migration, family formation, and socioeconomic inequality. Court rulings and legal judgments provide critical insights into how laws are interpreted and applied in ways that directly impact individuals and families. For example, divorce judgments contain extensive details on custody arrangements, property division, and financial obligations—information that is often missing or overly generalized in traditional survey data. Similarly, legal decisions on immigration cases, reproductive rights, and labor disputes reflect broader structural forces that shape demographic behaviors and disparities. Despite their richness, legal judgments remain an underutilized data source in demographic research. Their sheer volume, complexity, and unstructured nature pose significant challenges for systematic analysis. Unlike structured datasets, legal texts are lengthy, variable in format, and embedded with specialized legal language, making large-scale manual coding impractical. As a result, key demographic insights embedded in legal rulings often go unnoticed, limiting our ability to fully understand the social and economic forces driving demographic change.

Despite the rich demographic-relevant information embedded in legal judgments, traditional approaches to analyzing such texts face significant limitations. Manual coding and small-scale qualitative analyses, while useful for capturing nuanced interpretations, are inherently labor-intensive, time-consuming, and lack scalability. This makes them impractical for analyzing large volumes of legal documents systematically. Existing computational approaches—such as dictionary-based classification (Leahey et al., 2023; Velasco & Paxton, 2022), topic modeling and clustering (Duxbury, 2024; Greve et al., 2022; Wetts, 2023), and word vector representations with supervised learning (Best & Arseniev-Koehler, 2023; Kim & Askin, 2024; Nelson et al., 2023;

Wilmers & Zhang, 2022; Zhou, 2022)—are widely used in social science research but struggle with complex legal texts. These methods help identify key themes or variables but face computational and methodological challenges. Supervised learning, for instance, requires extensive labeled data and significant computational resources (Al Medawer, 2023). Moreover, legal rulings involve structured yet context-dependent language, with judges making intricate, case-specific determinations—such as nuanced custody arrangements or wealth divisions—that go beyond simple classification tasks. This complexity makes legal text analysis more challenging than typical natural language processing (NLP) applications, which often focus on sentiment analysis or document classification. These limitations are even greater in non-English contexts, where labeled legal datasets are scarce, and rule-based methods struggle with the complexity of judicial language (Michelson, 2019). For instance, Chinese divorce judgments contain intricate details on property disputes and custody decisions, requiring advanced NLP techniques for structured analysis. Similar challenges arise in other legal systems, where distinct traditions make it difficult to apply models trained on English-language corpora, leaving legal documents an underutilized resource in demographic research.

This study harnesses LLMs to analyze 3,602 randomly selected legal judgments from Zhejiang Province, China, drawn from a larger dataset of over 72,102 divorce judgments nationwide from 2009 to 2016, introducing a scalable and systematic approach to demographic research. Unlike traditional NLP methods that rely on predefined rules or extensive manual coding, fine-tuned LLMs, such as GPT-4o-mini, achieve higher accuracy in extracting nuanced demographic variables, including gender, custody claims, and property division. By leveraging deep neural networks and vast training corpora, these models excel in interpreting ambiguous language, detecting implicit social cues, and correcting human errors, offering a more robust

alternative to manual and rule-based classification methods. To assess their effectiveness, we compare ten computational strategies, from rule-based classification to fine-tuned transformer models, using a manually labeled gold-standard dataset. Our results show that fine-tuned decoder-only models achieve the best inference performance, with larger-scale models generally yielding better results. However, model performance depends not only on architecture but also on factors such as hyperparameter tuning, preprocessing techniques, and linguistic context. By demonstrating the scalability and efficiency of LLMs in processing large legal datasets, this study highlights their potential to reduce the resource intensity of traditional methods while enabling more comprehensive analyses of legal and policy-driven demographic trends.

This study is the first to apply LLMs to analyze Chinese divorce judgments for demographic and family research. Traditional NLP and rule-based methods have been used in legal text analysis but often struggle with the complexity of judicial language. For instance, Michelson (2019) used a rule-based approach for gender classification using the same data as ours and reported 70% missing data, highlighting its limitations in imputing demographic information. In contrast, our LLM approach accurately identifies gender with only 14% missing data, ensuring completeness and accuracy in demographic variable imputation. By leveraging LLMs, we demonstrate a scalable, accurate, and reliable method for extracting key demographic variables, providing new empirical insights into family dynamics and broader demographic processes.

Beyond these empirical contributions, our study introduces a methodological advancement by systematically comparing rule-based classification, traditional machine learning, and transformer-based models. Previous research has typically focused on individual LLM model types rather than comparing multiple approaches. For example, some studies have examined only transformer-based models (Chae & Davidson, 2023), while others have analyzed encoder-only

models with supervised learning (Wankmüller, 2024) or evaluated the zero-shot and few-shot capabilities of decoder-only models like GPT (Overos et al., 2024; Törnberg, 2024). However, a comprehensive comparison across different computational methods—including rule-based classification, traditional machine learning, and various LLM architectures—remains largely absent. By addressing these gaps, our work offers a framework for selecting the most effective text classification methods based on variable characteristics, text structure, and computational trade-offs.

More broadly, LLM-driven legal text analysis enables demographers to process vast, unstructured datasets at an unprecedented scale, overcoming the limitations of manual coding and survey-based research. While this study focuses on a sample of 3,602 cases to illustrate our approach, our method can efficiently code over 1.5 million legal judgments across all provinces in China from 2009 to 2016 within weeks, demonstrating its scalability and efficiency. The ability to analyze millions of legal documents opens new avenues for studying family law, socioeconomic disparities, and policy impacts, particularly in underexplored settings like many non-Western countries. As AI models continue to evolve, our findings encourage a reassessment of the role of expert coders in research, highlighting both the transformative potential and challenges of integrating LLMs into social science methodologies.

Data and Methods

The divorce cases analyzed in this study were sourced from the Zhejiang Provincial High Courts, covering rulings from 2009 to 2016 (Michelson, 2019). From 72,102 first-attempt divorce adjudications, we randomly selected 3,602 cases as a proof of concept for demonstrating the effectiveness of LLMs in legal text analysis. Given the dataset's scale—totaling 100 million

Chinese characters—manual coding would be impractical. Instead, we manually coded 5% of the dataset (3,602 cases) as a benchmark for validation.

Although divorce judgments follow a structured legal template with sections such as court details, litigants' statements, evidence, and final rulings, they are stored as unstructured HTML files encoded in GB18030, complicating data extraction. To provide further clarity, we include an example of a legal case in the appendix, along with English translations, offering a clearer sense of the structure and complexity of these documents.

We will compare manual coding results with ten computational strategies, including rule-based classification, traditional machine learning models such as Random Forest and LightGBM, and transformer-based models, including RoBERTa, GPT, and Qwen2.5. For all traditional machine learning and fine-tuned Transformer-based models, we used 3,002 examples for training and reserved 600 cases for testing. Our analysis focuses on three key variables: gender of the plaintiff and defendant, child custody claims, and housing property claims, as they provide critical insights into divorce adjudications, particularly regarding economic resource allocation and parental rights. Gender is coded as 0 when the plaintiff is female and the defendant is male, 1 when the plaintiff is male and the defendant is female, and 2 when gender is uncertain. Custody claims are categorized into five groups: 1 for cases where no claims were made, 2 when custody was awarded to the plaintiff, 3 when awarded to the defendant, 4 when shared custody was granted, and 5 when claims were made but not ruled upon. Housing claims are classified as 1 for cases with no claims, 2 when a claim was made and was awarded to the plaintiff, and 3 when a claim was made and was awarded to the defendant, 4 shared property ownership, and 5 when claims were made but not ruled upon.

Manual Coding

Coding gender, child custody claims, and housing property claims is complex, requiring an understanding of social cues, implicit language, and legal nuances. Gender identification is particularly challenging, as legal documents often omit direct references to protect litigants' privacy, forcing researchers to rely on indirect clues such as naming conventions or gendered roles—both of which can be unreliable when social norms deviate. In China, patriarchal naming conventions often provide useful clues for identifying family relationships, as children traditionally take their father's surname. However, this assumption can be unreliable when families deviate from tradition or when both parents share the same surname. Similarly, gendered roles—such as the expectation that husbands are primary breadwinners—can serve as helpful proxies in some cases but may lead to inaccuracies when social roles and economic responsibilities differ from traditional norms. While these indirect cues can aid classification, they require careful validation to ensure accuracy. Custody classification is complicated by ambiguous phrasing, with courts using indirect language like "the child will reside with" rather than explicitly assigning custody. This issue is further amplified in joint custody cases or when siblings are split between parents. Property disputes add another layer of uncertainty, as ownership claims may shift during proceedings, with assets initially classified as marital property later redefined as pre-marital. Joint ownership, extended family involvement, and shared property arrangements further complicate classification, requiring careful case-by-case analysis.

To ensure consistency and accuracy in the manual coding process, three native Chinese-speaking PhD students with expertise in divorce law were trained to code the data. As all coders were female, the male authors also reviewed the coding to help mitigate potential gender biases in interpretation. The coders cross-checked each other's work, and the final dataset was further

validated by the authors to maintain coding consistency. In cases where explicit gender mentions were absent, we provided guidelines (Table 1) to help coders incorporate their understanding of social cues while following a systematic approach. These guidelines ranked criteria from most to least reliable, prioritizing direct gender mentions, physiological characteristics, and social roles when inferring gender. Coders applied the highest-ranking applicable criterion or used a combination of indicators when necessary to ensure accurate classification. For custody and property claims, we did not establish additional detailed coding rules, as these were generally explicitly stated in the judgment documents. The manually coded dataset serves as the benchmark for evaluating our automated methods, ensuring a reliable reference for assessing model performance.

Rule-Based Classification

To automate the extraction of key variables, we developed a rule-based method using fuzzy matching and regular expressions built on a curated keyword corpus. Rule-based methods work by applying predefined linguistic rules—such as pattern matching, word relationships, and syntactic structures—to systematically extract information from unstructured text. Unlike machine learning models, which rely on statistical patterns, rule-based methods explicitly define how specific words or phrases should be identified and categorized.

To account for variations in language and address imbalanced distributions in smaller categories, we adopted an iterative trial-and-error approach to refine patterns indicative of each variable. For custody claims, we identified judicial decisions based on keywords such as plaintiff (原告), defendant (被告), custody (抚养), and live with (随 or 生活) while limiting our analysis to single-child cases, as joint custody rulings are often more complex. Common judicial phrases

for custody determinations include: “The girl Li XX will be raised by the defendant” (女孩李XX 由被告抚养) and “The eldest daughter, Chen XX, will be raised by the plaintiff” (婚生长女陈某某由原告抚养).

For housing property claims, we identified relevant cases using keywords such as housing (房), ownership (所有), plaintiff (原告), and defendant (被告) while excluding unrelated terms like renting (租房). To determine judicial rulings, we searched for key phrases such as “give up” (放弃) and “ownership” (产权). Typical judicial statements include: “The house at XX belongs to the plaintiff” (XX 房归原告所有) and “The plaintiff gives up their rightful share of the house at XX” (XX 房原告放弃应得份额).

Unlike custody and housing property claims, gender classification is more complex as it relies on implicit linguistic cues shaped by social and cultural norms, particularly in first-person litigant narratives. The variability of gendered expressions made keyword-based rules unreliable, so we primarily used litigants’ names, leveraging China’s patriarchal naming conventions to infer gender. When available, we compared children’s surnames with their parents’, as children in China traditionally take their father’s surname. To extract surnames, we identified key expressions related to birth (出生), child (孩子), son (儿子), and daughter (女儿), which frequently appear in discussions of custody and parental responsibilities. Once a child’s full name was detected, we extracted the surname and compared it against the litigants’ names. If a match was found, we inferred the father’s identity based on naming conventions. Recognizing exceptions in surname inheritance, we incorporated additional checks. When both parents shared a surname or the child’s surname differed from both litigants, we examined contextual references—such as parental roles and financial responsibilities—to improve classification

accuracy. This multi-step approach helped reduce misclassification in cases where traditional naming patterns did not apply.

Supervised Machine Learning

Supervised machine learning (SML) algorithms are widely used for text classification, offering a scalable alternative to rule-based methods, which rely on predefined linguistic patterns and often struggle with ambiguous or novel expressions (Nelson et al., 2021). In this study, we apply LightGBM, a gradient boosting decision tree (GBDT) model, known for efficiently handling high-dimensional feature spaces and large datasets while maintaining accuracy comparable to conventional GBDT models (Ke et al., 2017). Unlike rule-based methods, which require manually crafted rules that may not generalize well across diverse texts, LightGBM learns patterns from data, constructing multiple decision trees to classify text based on extracted features. Its ability to handle sparse data and complex relationships in text makes it particularly effective for legal document classification, where wording and structure can vary significantly. Additionally, LightGBM offers faster training times compared to deep learning models, making it a practical choice for large-scale text analysis.

To generate the feature matrix, we used OpenAI's pre-trained embedding model, "text-embedding-3-large," which produces 3,072-dimensional vector representations at the document level. This model was selected to accommodate the length and semantic complexity of our legal texts. Text preprocessing involved extracting the main body of each document using regular expressions, starting from plaintiff claims (e.g., "原告起诉称") and ending with concluding legal statements (e.g., "如不服本判决"). Any text exceeding 8,100 tokens was truncated to fit within processing limits. Since embedding models require minimal text preprocessing (Wankmüller,

2024), we fed the processed text directly into the embedding model to obtain vector representations. The embedding vectors and their corresponding labels were then used to train LightGBM using the training dataset and evaluate its performance on the test set. All experiments were conducted in Python on Google Colab without GPU acceleration. We implemented LightGBM via the lgbm library, optimizing hyperparameters through 5-fold cross-validation to enhance model performance.

Transformer-based Models

Transformer-based models take a different approach from traditional machine learning models like LightGBM, which relies on precomputed embeddings and structured feature representations. While LightGBM is computationally efficient, it may struggle with nuanced legal language and implicit relationships between words. In contrast, Transformers dynamically learn contextual relationships, making them better suited for complex linguistic structures found in legal judgments (Chae & Davidson, 2023; Wankmüller, 2024).

Within the Transformer architecture, there are two main types of models: encoder-only and decoder-only. Encoder-only models, such as RoBERTa, process text bidirectionally, analyzing a token in the context of both preceding and following words. This structure makes them particularly effective for classification tasks requiring full-text comprehension. Decoder-only models, such as GPT and Qwen2.5, operate autoregressively, meaning each token attends only to previous tokens. While this is ideal for text generation, it may be less effective for classification tasks that require a holistic view of the text. In this study, we compare encoder-only models (RoBERTa) and decoder-only models (GPT, Qwen2.5) to evaluate their effectiveness in legal text classification. Given that legal judgments contain both explicit classifications (e.g., custody rulings) and implicit

indicators (e.g., gender inference from social cues), theoretically, encoder-based models may be better suited for tasks requiring comprehensive text understanding.

We fine-tuned the pretrained RoBERTa model, Chinese-roberta-wwm-ext (Cui et al., 2021), which was trained with whole word masking (WWM) on an extended dataset of 5.4 billion tokens from Wikipedia and other publicly available corpora. The model was accessed via Hugging Face and fine-tuned on Google Colab. Text preprocessing involved extracting the main body of each judgment document using regular expressions, removing standard opening and closing phrases. Since RoBERTa has a 512-token limit, we applied a sliding window approach, chunking documents into 510-token segments with 50-token overlaps to retain contextual coherence. The Hugging Face Trainer framework was used for fine-tuning with learning rate = 4e-5, batch size = 16, 3 epochs, and weight decay = 0.01, conducted on an NVIDIA A100 GPU (40GB) with PyTorch CUDA acceleration. For inference, we trained separate models for each variable, applying the same chunking strategy and aggregating logits across all chunks using a weighted average approach to determine the final classification.

We evaluated a range of GPT models, starting with zero-shot learning using GPT-3.5-turbo via the OpenAI API, where the model classifies categories without prior exposure to labeled examples (Radford et al., 2019; Yin et al., 2019). We then applied few-shot learning, incorporating labeled examples within prompts to improve accuracy. Expanding this approach, we conducted inference using GPT-3.5, GPT-4, and GPT-4o, with model sizes ranging from 175 billion to 1.7 trillion parameters (Chae & Davidson, 2023). To improve consistency and accuracy, we fine-tuned GPT-4o-mini, a more efficient model, as recommended by Chae and Davidson (2023) for balancing performance and computational cost. Unlike BERT models, where input text is paired with labels, fine-tuning GPT models required structuring prompts in JSON format to align with

their autoregressive decoding structure. We fine-tuned the model in progressive stages, training on 200 and 800 cases before scaling up to 3,002 cases, with the remaining 600 cases reserved for testing, while keeping all hyperparameters at default values.

Using a commercial API for academic research raises concerns about transparency, replicability, and cost, prompting us to experiment with an open-source LLM, Qwen2.5-32B-Instruct. This model was selected for its extensive training on Chinese corpora and compatibility with our high-performance computing (HPC) cluster. We set up a virtual environment, installed dependencies via pip, and prepared labeled data and prompts in ShareGPT format. Fine-tuning was conducted using DeepSpeed for efficient distributed training, applying LoRA-based fine-tuning to optimize query and value projection layers (q_proj, v_proj) while minimizing computational overhead. The model was trained with a 3,072-token sequence length, a batch size of 1 per device, gradient accumulation over 16 steps, and a learning rate of 5e-6 with a cosine scheduler over three epochs. To improve efficiency, we leveraged bf16 precision, FlashAttention, and DeepSpeed ZeRO-3 for optimized memory distribution across four NVIDIA A100 GPUs, each paired with two CPU cores, and submitted all tasks via SLURM. For inference, we loaded the fine-tuned model using Hugging Face's Transformers library, employing AutoModelForCausalLM and AutoTokenizer to execute inference with GPU acceleration via PyTorch CUDA.

Evaluation Metrics

To evaluate model performance in replicating human coding, we used a confusion matrix, focusing on macro-average F1 and weighted-average F1 scores, both commonly used in multiclass classification. The macro-average F1 score calculates the F1 score for each class independently

and then averages them equally across all categories, regardless of their frequency. This makes it particularly useful for evaluating model performance on underrepresented classes, ensuring that smaller categories are not overshadowed by more frequent ones. However, it may not fully reflect overall model accuracy if the dataset is highly skewed toward certain labels. In contrast, the weighted-average F1 score accounts for class imbalance by weighting each F1 score based on the proportion of instances in that class. This ensures that performance on dominant categories contributes more heavily to the final score, providing a more realistic measure of overall classification accuracy in imbalanced datasets.

Findings

Figure 1 illustrates the percentage distribution of gender, custody, and housing property claims from manual coding. Gender distribution shows a clear imbalance, with female plaintiffs and male defendants accounting for 55.3% of cases, while male plaintiffs and female defendants represent 30.6%. A notable 14.1% of cases have uncertain gender information, indicating ambiguous references in legal documents. In terms of custody, 33.8% of cases do not include custody claims, while among cases with claims, plaintiff custody (12.6%) is more common than defendant custody (6.0%), and shared custody is the least frequent at just 1.2%. A significant 46.4% of cases lack an explicit custody ruling despite custody claims. In some instances, judges may decide not to rule on custody within the divorce judgment, instead stating that such matters should be resolved in a separate legal proceeding. Housing property claims are even rarer, with 89.6% of cases containing no such claims. Among the 10.4% of cases where property claims were made, 8.2% lacked a judicial ruling, while only 0.7% granted property to the plaintiff, 0.8% to the defendant, and 0.6% involved shared property ownership. These findings suggest that custody and property disputes are

often settled outside of formal court rulings and that gender asymmetry in plaintiffs may reflect broader social or legal factors influencing divorce proceedings.

[Figure 1 About Here]

Performance of coding these variables across approaches is shown in Table 2. Overall, fine-tuned Transformer-based models consistently outperformed rule-based classification and traditional machine learning methods, as shown by the Macro and Weighted F1 scores. Among these models, GPT-4o-mini demonstrated high inference stability across variables, despite differences in context dependency and class distribution. It achieved F1 scores of 0.88 (macro) and 0.91 (weighted) for gender, 0.73 (macro) and 0.93 (weighted) for housing, and 0.93 for both metrics in custody classification, highlighting its ability to handle complex linguistic structures and imbalanced distributions in legal texts. For few-shot learning, we tested 360 labeled cases and found that adding classification examples to the prompt improved the weighted F1 score from 0.57 to 0.60. However, switching the prompt language from English to Chinese while keeping examples in Chinese lowered the score to 0.56, highlighting the sensitivity of performance to prompt design. Testing GPT-3.5, GPT-4o-mini, GPT-4, and GPT-4o on the same prompts confirmed the scaling law, with larger models consistently performing better. GPT-4o achieved the highest F1 score of 0.80, reinforcing that larger models excel in few-shot classification tasks.

Among other decoder-only models, Qwen2.5-32B-Instruct lagged behind GPT-4o-mini in all three tasks—gender (macro F1: 0.27), property (macro F1: 0.56), and custody (macro F1: 0.52). While scaling laws suggest that larger models tend to perform better, model size alone does not fully explain the performance gap. Fine-tuning efficiency, influenced by optimization techniques such as bf16 precision and LoRA, played a crucial role in improving results, especially given HPC resource constraints. Notably, fine-tuned decoder-only models were less sensitive to class

imbalance, as seen in their smaller gap between macro and weighted F1 scores. This suggests that when certain categories contain very few labeled examples—such as housing property classifications, where "granted to plaintiff" (0.7%), "granted to defendant" (0.8%), and "shared property" (0.6%) each had fewer than 30 instances—decoder-only models remain robust in handling rare-category predictions.

Encoder-only models, such as RoBERTa, are widely recognized for their effectiveness in text classification, and our results support this. Fine-tuned RoBERTa performed well in gender classification (macro F1: 0.82, weighted F1: 0.84) and custody classification (macro F1: 0.73, weighted F1: 0.87), though its performance on housing property claims was lower. This drop is largely due to the scarcity of labeled examples, as minority categories collectively accounted for less than 2.1% of the dataset. Notably, in gender classification, which depends heavily on semantic context, cultural cues, and implicit linguistic markers, RoBERTa ranked second only to GPT-4o-mini. Given its simpler architecture, smaller pretraining corpus, and lack of prompt engineering, RoBERTa offers a cost-efficient alternative for consistent labeling, easily deployable via Hugging Face pipelines on local machines or Google Colab with GPU acceleration. However, a key limitation of BERT-based models is their 512-token limit, requiring document chunking and logit aggregation during inference.

Traditional machine learning models, such as LightGBM, performed significantly worse than fine-tuned RoBERTa and GPT-4o-mini. The largest performance gap was observed in gender classification, suggesting that the embedding model used in LightGBM failed to capture key linguistic and structural patterns in legal texts. Additionally, the large difference between macro and weighted F1 scores in housing and custody classification highlights a major limitation of decision tree-based models: they require a sufficient number of examples in all categories to

generalize well. Given the highly imbalanced nature of our dataset, LightGBM overfit to majority categories, struggling with minority-class predictions—a common challenge for traditional machine learning approaches in legal text classification.

Rule-based classification depends heavily on capturing linguistic variations and is highly sensitive to class imbalances. In housing property classification, its macro and weighted F1 scores were comparable to those of fine-tuned GPT-4o-mini, demonstrating its potential when linguistic patterns are well-defined. However, custody classification yielded suboptimal results (macro F1: 0.35, weighted F1: 0.33), suggesting that predefined rules struggled to generalize across cases. Rule-based gender classification was even more challenging, as gender expression in first-person narratives is deeply embedded in cultural and social contexts, making it difficult to establish fixed rules.

Table 3 compares the time required and computing resources for different text classification methods. Rule-based classification is the fastest, taking only three minutes per variable on a personal computer, while LightGBM requires 15 minutes for training and prediction on similar hardware. Fine-tuned Transformer models demand significantly more resources, with RoBERTa taking 30 minutes on a single A100 GPU, GPT-4o-mini requiring 40 minutes via a commercial API, and Qwen2.5-32B-Instruct taking the longest at 80 minutes, utilizing four A100 GPUs on an HPC node. This highlights the trade-off between efficiency and computational demand, with larger models requiring more time and resources for fine-tuning and inference.

Discussion and Implications

Our findings suggest that fine-tuned transformer-based models can produce consistent and accurate results, particularly for text classification tasks that are highly dependent on contextual

interpretation. Given limited computing resources, fine-tuning LLMs via commercial APIs on a personal computer presents one of the most accessible approaches. However, if replication, data privacy, or transparency are key concerns, fine-tuning BERT-based models can serve as an efficient alternative, though researchers must consider maximum token length limitations. Another option is fine-tuning open-source LLMs, which offers greater flexibility and control but typically requires more computational resources than a personal computer, as well as a higher level of technical expertise, including proficiency in command-line scripting and vRAM optimization. For tasks without rare categories, vectorized representations from embedding models combined with conventional machine learning can provide a fast, cost-effective alternative to LLM fine-tuning. When the sample size is small, zero-shot or few-shot learning can be employed for annotation. Alternatively, rule-based methods can be particularly effective when classification depends primarily on key terms with limited variation rather than broader semantic context.

While each method has its optimal application, manual coding remains essential for validation. Given GPT-4o-mini's strong performance on the test dataset, we applied the fine-tuned model to infer 69,100 cases, excluding the training set. Running batch inference via the API took approximately seven hours and cost $12.60, demonstrating a cost-effective and time-efficient approach. However, this efficiency depends on the availability of well-labeled training data, as the model's accuracy is inherently tied to annotation quality. While computational methods can replicate human coding and scale analysis without concerns of fatigue, high-quality labeled data remains a critical prerequisite for reliable results.

The inconsistencies between AI and human coding reveal key challenges in gender classification and the broader question of what constitutes ground truth in text annotation. In our review of 100 gender-uncertain cases, 56 lacked identifiable gender markers, aligning with the

model's inference. However, in 29 cases involving harassment, violence, threats, gambling, or prostitution, human coders—following the codebook guidelines—were more likely to classify the agent of action as male, while GPT-4o-mini labeled these cases as gender uncertain, suggesting a more conservative approach that avoids assumptions when explicit markers are absent. This discrepancy raises important questions: Should gender classification incorporate statistical patterns of behavior, such as the higher likelihood of men engaging in gambling or domestic violence? Can first-person narrative styles, often linked to masculine or feminine linguistic patterns, serve as reliable classification cues? Or should cases relying on implicit social cues always default to gender uncertain?

A more effective approach would involve quantifying uncertainty by using probability distributions rather than rigid categorical labels. For example, if the probability of a person being male exceeds 0.8, they could be classified as male, whereas lower probabilities might warrant an uncertain label, making classification decisions more transparent and better reflecting cases where gender inference depends on contextual rather than explicit markers. To improve AI-human alignment, models need to incorporate probabilistic reasoning, weighing contextual and behavioral cues in ways that align with research objectives. The differences between AI and human coding are not necessarily errors but rather distinct approaches to interpreting social norms—humans may rely more on statistical associations or implicit biases, while GPT-4o-mini appears to take a neutral stance, avoiding gender assumptions in ambiguous cases. By integrating probabilistic classification thresholds, AI systems can move beyond binary coding, creating a more flexible, interpretable, and empirically grounded approach to text annotation.

Our study has four key limitations. First, rare categories in LightGBM and RoBERTa could be better handled using oversampling techniques like SMOTE or adjusting class weights during

fine-tuning to improve macro F1 scores. Second, while we focused on Qwen2.5-32B-Instruct, other advanced open-source Chinese LLMs could be explored. Computational constraints limited us to parameter-efficient fine-tuning (LoRA) on a 32-billion-parameter model, but full fine-tuning on a larger model may enhance accuracy. Third, our analysis was limited to gender, housing property division, and custody claims, but divorce judgments contain richer contextual data on marital behaviors, extended family dynamics, and marriage history, offering opportunities for future research. Lastly, while we prioritized classification accuracy (F1 scores), interpretability remains a concern, particularly for socially embedded variables like gender. To improve transparency, interpretable modeling tools could be integrated to highlight key linguistic cues, or a customized approach for each variable—such as mapping first-person narratives to gendered anchor words using pretrained embeddings—could enhance explainability in socially complex classifications.

This research highlights how AI-assisted coding can transform family and demographic studies by extracting structured variables from large-scale legal and textual data, supplementing key information often missing in surveys. Beyond automating classification, LLMs offer a tool for critically reassessing measurement approaches, particularly for demographic characteristics deeply embedded in sociocultural contexts. As researchers refine computational methods, they can extend AI-driven text analysis to capture complex family dynamics, analyze biases in classification, and even construct counterfactuals for studying social change. Future applications of LLM-based classification in causal inference hold the potential to revolutionize how family demographers explore evolving legal, social, and economic trends at scale.

References

Best, R.K., & Arseniev-Koehler, A. (2023). The Stigma of Diseases: Unequal Burden, Uneven Decline. *American Sociological Review*, 88, 938-969.

Chae, Y., & Davidson, T. (2023). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning. SocArXiv.

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3504-3514.

Duxbury, S.W. (2024). Collaborating on the Carceral State: Political Elite Polarization and the Expansion of Federal Crime Legislation Networks, 1979 to 2005. *American Sociological Review*, 89, 650-683.

Greve, H.R., Rao, H., Vicinanza, P., & Zhou, E.Y. (2022). Online Conspiracy Groups: Micro-Bloggers, Bots, and Coronavirus Conspiracy Talk on Twitter. *American Sociological Review*, 87, 919-949.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kim, K., & Askin, N. (2024). Feature-Based Structures of Opportunity: Genre Innovation in the American Popular Music Industry, 1958 to 2016. *American Sociological Review*, 89, 542-583.

Leahey, E., Lee, J., & Funk, R.J. (2023). What Types of Novelty Are Most Disruptive? *American Sociological Review*, 88, 562-597.

Nelson, L.K., Brewer, A., Mueller, A.S., O'Connor, D.M., Dayal, A., & Arora, V.M. (2023). Taking the Time: The Implications of Workplace Assessment for Organizational Gender Inequality. *American Sociological Review,* 88, 627-655.

Nelson, L.K., Burk, D., Knudsen, M., & McCall, L. (2021). The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research,* 50, 202-237.

Overos, H.D., Hlatky, R., Pathak, O., Goers, H., Gouws-Dewar, J., Smith, K., et al. (2024). Coding with the machines: machine-assisted coding of rare event data. *PNAS Nexus,* 3.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog,* 1, 9.

Törnberg, P. (2024). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review,* 0, 08944393241286471.

Velasco, K., & Paxton, P. (2022). Deconstructed and Constructive Logics: Explaining Inclusive Language Change in Queer Nonprofits, 1998–2016. *American Journal of Sociology,* 127, 1267-1310.

Wankmüller, S. (2024). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological Methods & Research,* 53, 1676-1752.

Wetts, R. (2023). Money and Meaning in the Climate Change Debate: Organizational Power, Cultural Resonance, and the Shaping of American Media Discourse. *American Journal of Sociology,* 129, 384-438.

Wilmers, N., & Zhang, L. (2022). Values and Inequality: Prosocial Jobs and the College Wage Premium. *American Sociological Review,* 87, 415-442.

Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Zhou, D. (2022). The Elements of Cultural Power: Novelty, Emotion, Status, and Cultural Capital. *American Sociological Review*, 87, 750-781.

Tables

Table 1 Gender Classification Criteria for Manual Coding

Order	Criteria	Explanation	Example
1	Direct gender mentions	Direct Gender Mentions: Identifies explicit gender terms, such as '男' (male) and '女' (female), referring to the plaintiff or defendant.	Plaintiff: Fan, Male (原告：樊甲，男); Defendant: Li, Female (被告：李某某，女).
2	Physiological gender characteristics	Recognizes gender-specific physical conditions or reproductive events, such as pregnancy or miscarriage.	"Plaintiff had multiple abortions" (原告曾多次做过人工流产手术) indicates a female plaintiff.
3	Social roles related to gender	Infers gender based on cultural norms, customs, and terminology used in marriage or intimate relationships.	"Plaintiff demands the return of the dowry from the defendant" (要求被告返还嫁妆) suggests the defendant is male, as dowries are traditionally given to the groom's family.
4	Child information	Determines parental gender by analyzing children's surnames, which typically follow the paternal line.	"Gave birth to a daughter whose surname is Jin" (生育一女名金丙). If '金' matches the surname of one party, that party is likely the father (male).
5	Marital conflicts	Uses gender-associated behaviors in marital disputes, such as domestic violence or gambling, which are more commonly linked to males.	"Defendant neglected family duties, repeatedly deceived and hurt the plaintiff, and gambled illegally" (被告未尽家庭责任,多次欺骗伤害原告，非法赌博借债), suggesting the defendant is male.
6	Gender-indicative name features	Identifies gender based on name characteristics traditionally associated with males or females.	"Defendant Li Zhao Di" (被告李招娣) suggests a female defendant, as '招娣' historically indicates a preference for a future male sibling.

Table 2. Comparing Coding Performance Across Methods

	Child Custody			Housing Property			Gender		
	Accuracy	Weighted F1	Macro F1	Accuracy	Weighted F1	Macro F1	Accuracy	Weighted F1	Macro F1
Rule-based	0.31	0.35	0.33	0.9	0.91	0.72	0.56	0.6	0.52
Random Forest	0.76	0.73	0.46	0.9	0.85	0.19	0.59	0.48	0.37
LightGBM	0.80	0.78	0.51	0.9	0.86	0.24	0.66	0.62	0.56
Fine-tuned RoBERTa	0.87	0.87	0.73	0.91	0.90	0.38	0.84	0.84	0.82
Zero or Few-shots:									
GPT-3.5							0.60	0.58	0.48
GPT-4							0.73	0.72	0.66
GPT-4o							0.71	0.73	0.68
GPT-4o-mini							0.46	0.46	0.43
Fine-tuned Qwen2.5							0.57	0.43	0.28
Fine-tuned GPT-4o-mini	0.93	0.93	0.93	0.93	0.93	0.73	0.92	0.91	0.88

Note: $F1_{macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times TP_i}{2\times TP_i+FP_i+FN_i}$, $F1_{weighted} = \sum_{i=1}^{N}\frac{n_i}{\sum_{j=1}^{N} n_j}\frac{2\times TP_i}{2\times TP_i+FP_i+FN_i}$, where TP_i is true positives for class *i*, FP_i is false positive for class *i*, FN_i refers to false negatives for class *i*, *N* indicates total number of classes, n_i points to number of true instances of class *i*.

Table 3 Time and Cost

Method	**Time Required per Variable**	**Computing Resources**
Rule-based	Classification: 3 minutes	Personal computer
LightGBM	Training and Prediction: 15 minutes	Personal computer
Fine-tuned RoBERTa	Fine-tuning and inference: 30 minutes	1 A100 GPU on Google Colab
Fine-tuned GPT-4o-mini	Fine-tuning and inference: 40 minutes	Commercial API
Fine-tuned Qwen2.5-32B-Instruct	Fine-tuning and inference: 80 minutes	4 A100 GPUs on a single HPC node

Notes:
The time and resources spent on manual annotation are not included in the table.
The time required for writing scripts and tuning configurations is not considered.
The queueing time for computing resources on HPC is not included.
The time required for document-level embedding when using LightGBM methods is not included.

Figures

Figure 1. Distribution of Key Variables