Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 18 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i18.58071

PAPER

Real-Time Legal Advisory System for Mobile Computing Devices via Deep Learning

Shuo Cheng (✉)

Law School, Guangxi University for Nationalities, Nanning, China

chengshuo777088@163.com

ABSTRACT

With the widespread adoption of mobile internet technologies and a growing public awareness of legal rights, demand for real-time and accurate legal advice in mobile environments has increased significantly. Traditional legal services are constrained by uneven resource distribution and delayed response times. Existing legal advisory systems, which often rely on static knowledge bases or simple rule-matching techniques, have demonstrated notable limitations in understanding personalized user needs, processing multimodal inputs, and adapting to mobile devices. Collaborative filtering methods require large-scale annotated datasets and are generally inadequate in capturing the domain-specific semantics of legal texts. Recurrent neural networks (RNNs) and other models impose computational burdens that hinder real-time responsiveness on mobile platforms. Moreover, they lack mechanisms for high-order feature integration and key information extraction in user-legal knowledge interactions. To address these challenges, a real-time legal advisory model based on an attention-compressed interactive network was proposed. By extracting interaction features between user consultation texts and legal knowledge units and integrating an attention mechanism to filter critical semantic information, a lightweight compressed interaction network was designed to enable high-order feature fusion while remaining suitable for devices with limited computational capacity. A score prediction module was incorporated to quantify the relevance of advisory responses, forming an end-to-end recommendation system. This approach overcomes dual bottlenecks in semantic modeling and device adaptability that hinder traditional models, providing a technical solution for generating efficient legal advice in mobile settings. The findings offer practical implications for deploying intelligent legal services on edge devices and contribute to the broader development of accessible legal infrastructure.

KEYWORDS

deep learning, mobile computing devices, real-time legal advisory system, attention-compressed interactive network, lightweight model

Cheng, S. (2025). Real-Time Legal Advisory System for Mobile Computing Devices via Deep Learning. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(18), pp. 77–91. https://doi.org/10.3991/ijim.v19i18.58071

Article submitted 2025-04-29. Revision uploaded 2025-07-09. Final acceptance 2025-07-20.

1 INTRODUCTION

With the proliferation of mobile Internet technologies [1–3] and the advancement of public legal awareness [4, 5], an exponential increase has been observed in the demand for real-time and accurate legal advice in mobile contexts. In daily life, legal issues such as contract review, dispute resolution, and rights protection frequently arise [6–9]. However, traditional legal services are hindered by uneven resource distribution, prolonged response times, and high professional entry barriers [10, 11], rendering them inadequate for addressing fragmented and real-time consultation needs in mobile scenarios. Due to their portability and widespread adoption, mobile computing devices have emerged as a critical interface between users and legal resources [12, 13]. Nevertheless, existing legal advisory systems predominantly rely on static knowledge bases or rudimentary rule-matching approaches, which lack the capacity for deep understanding of personalized user needs and dynamic interaction. These systems are particularly ineffective in processing complex case descriptions and multimodal inputs, where semantic alignment and knowledge recommendation must be executed efficiently. Against this backdrop, the development of a real-time legal advisory system tailored for mobile devices through deep learning methodologies is a pivotal challenge in enhancing the accessibility of public legal services.

Current research on intelligent recommendation in the legal domain has primarily employed conventional machine learning techniques or baseline neural network models. Some studies [14–16] have leveraged traditional recommendation algorithms to analyze users' historical consultation records; however, these methods are heavily dependent on large-scale annotated datasets and are generally incapable of capturing the specialized semantics inherent in legal texts. Other studies [17–19] have explored the application of artificial intelligence in processing legal consultation texts but have encountered limitations related to computational constraints of mobile devices. High model complexity has resulted in inadequate real-time responsiveness. Furthermore, traditional models [20, 21] often fail to integrate high-order features and filter key information effectively when modeling interactions between user input and legal knowledge units. For instance, vague expressions or polysemous terms in user queries are rarely addressed with targeted attention mechanisms, leading to reduced recommendation accuracy. More critically, limited consideration has been given to the hardware constraints of mobile devices, and significant deficiencies persist in model light-weighting and local deployment strategies. As a result, existing systems struggle to balance computational efficiency with advisory effectiveness.

Focusing on the demand for real-time legal advisory services on mobile computing devices, an intelligent model based on an attention-compressed mobile interactive network was proposed in this study. The model architecture comprises four core modules. First, interaction features between user consultation texts and legal knowledge units were extracted, with critical semantic information identified through an attention mechanism. Second, a lightweight compressed interactive network was designed to achieve high-order feature fusion between user input and legal knowledge, thereby addressing the computational limitations of mobile devices. Third, a score prediction module was incorporated to quantify the relevance of legal recommendations, enabling dynamic ranking and precise suggestion generation. Finally, an end-to-end recommendation system was constructed to

support the intelligent generation of legal knowledge units and real-time feedback. In contrast to existing research, the primary innovation of this work lies in the integration of an attention mechanism with a compressed interactive network. This approach preserves the semantic modeling capacity of deep learning for legal texts while significantly reducing model complexity through architectural optimization, making it suitable for deployment on resource-constrained mobile platforms. The findings offer not only a technically efficient solution for real-time legal advisory systems but also a methodological reference for legal text processing and lightweight model design. These contributions are expected to facilitate the transition of intelligent legal services from cloud-based frameworks to edge-side deployment, thereby advancing the implementation of accessible and inclusive legal infrastructure.

2 REAL-TIME LEGAL ADVISORY MODEL BASED ON AN ATTENTION-COMPRESSED MOBILE INTERACTIVE NETWORK

2.1 Model architecture

The foundational architecture of the real-time legal advisory model is centered on the dynamic interaction between user needs and legal knowledge, implemented through a four-layer modular design that enables an end-to-end pipeline from text processing to intelligent recommendation. Initially, legal consultation texts submitted by users and the corresponding legal knowledge units were vectorized and structured into feature representations, which were subsequently input into a compressed mobile interactive network specifically designed for resource-constrained mobile devices. Through multilayer nonlinear transformations, the network captured high-order semantic associations between user queries and legal knowledge units—for example, mapping critical legal elements expressed in user language to their corresponding legal concepts. An attention mechanism was employed to assign semantic weights to core components within the interaction texts, prioritizing ambiguous expressions, polysemous terms, and complex case details while filtering irrelevant or redundant content. This process resulted in the generation of mobile-interactive textual features that preserve essential semantic information.

Upon completion of core feature extraction, feature fusion was performed using activation functions to integrate mobile-interactive textual features with historical scoring data reflecting prior recommendation relevance. This cross-modal fusion yielded a final feature vector encapsulating both personalized user intent and the contextual attributes of legal knowledge. The user-side features emphasize deep semantic representation of consultation intent, while the legal knowledge unit features highlight the applicable legal domain and professional scope of the knowledge. The fused representation was then passed through a dense network to predict the relevance score between user input and legal knowledge units. A series of nonlinear transformations in fully connected layers were used to quantify the degree of alignment, providing a numeric basis for recommendation ranking and ensuring highly relevant legal suggestions are delivered in real time. Of particular note, the legal knowledge unit features are treated as a re-constructable component, dynamically optimized through a language model estimation mechanism. This enables

continuous refinement of knowledge representation. The basic model architecture is illustrated in Figure 1.

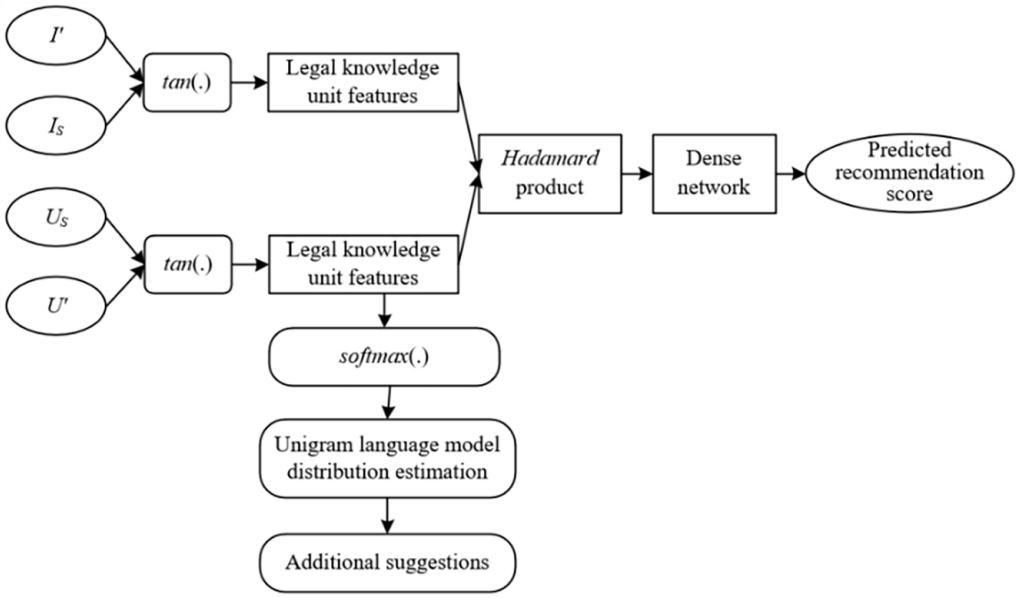

Fig. 1. Foundational architecture of the real-time legal advisory model

2.2 Mobile-interactive text feature extraction

The set of user-generated interactive texts on mobile devices is defined as $W_I = \{W_1, W_2, ..., W_j\}$, where W_j denotes the j-th interactive text, and s represents the total number of user interactions. These texts comprise multi-turn dialog inputs submitted to the real-time legal advisory system, including case descriptions, historical consultation records, and supplementary legal inquiries. To accommodate the computational constraints of mobile devices, lightweight text encoding techniques were adopted to transform each interactive text into a low-dimensional dense vector. This transformation preserves essential legal semantic information while reducing data dimensionality. The resulting encoded vector sequences were input into the compressed mobile interactive network, which was structurally optimized for limited memory and computational capacity. This optimization involved parameter sharing and inter-layer connectivity reduction to ensure efficient on-device operation. Assuming the word vector at the c-th position of the k-th interactive text is denoted as $W_{k,c}$, and the total number of words in that text is c, then the concatenated word vectors of each interactive text can be represented as:

$$W_k = \{W_{k,1}, W_{k,2}, \ldots, W_{k,c}\} \tag{1}$$

The encoded k-th interactive text W_k was then passed through the compressed mobile interactive network, where high-order feature representations were modeled via multiple layers of nonlinear transformations. Lightweight convolutional modules or self-attention mechanisms were embedded within the network to identify relationships among domain-specific legal terms and to capture the logical structure

inherent in user queries. The output feature matrix of each network layer, denoted as $W^{j}_{y,*}$, represents semantic abstractions at various hierarchical levels, encompassing both lexical-level legal concepts and sentence-level logical relationships. Let W^{j} denote the feature matrix at the j-th layer, with the total number of layers represented by j. The number of feature vectors in the $(j-1)$-th layer is denoted G_{j-1}, and the parameter matrix of the u-th feature vector is represented for $1 \le y \le G_j$. The Hadamard product is denoted by $\cdot$. Thus, the complete set of features in the y-th row vector of the feature matrix at layer j, denoted as $W^{j}_{y,*}$, can be expressed as:

$$W^{j}_{y,*} = \sum_{u=1}^{G_{j-1}} \sum_{k=1}^{s} Q^{j,y}_{uk} \left(W^{j-1}_{u,*} \cdot W^{0}_{k,*} \right) \tag{2}$$

To process the output of each hidden layer within the compressed mobile interactive network, a computationally efficient sum pooling operation was applied to transform high-dimensional feature maps into low-dimensional pooled vectors, denoted as $T^{j}_{G_j}$. The primary objective of this pooling operation is to reduce feature dimensionality while retaining the overall semantic representation, thereby mitigating the computational latency typically induced by high-dimensional data on mobile devices. For instance, in user-submitted legal queries involving detailed narratives on "traffic accident liability determination," the pooling operation enables concentrated representation of core semantics such as "responsibility allocation" and "damage compensation," while filtering out irrelevant content. This lightweight processing strategy is well aligned with the real-time interaction requirements of mobile devices, ensuring rapid feature extraction on edge platforms. Let G_j denote the total number of pooled vectors, and Z represent the dimensionality of the embedding feature space. The word vector located at the k-th position of the G_j-th row in the j-th layer feature matrix is denoted as $W^{j}_{G_j,k}$. Then $T^{j}_{G_j}$ is formally defined as:

$$T^{j}_{G_j} = \sum_{k=1}^{Z} W^{j}_{G_j,k} \tag{3}$$

The output layer of the compressed mobile interactive network integrates the pooled vectors from all hidden layers with the original encoded information to generate the preliminary mobile-interactive text features. This output layer is implemented as a lightweight, fully connected layer with pruned parameters to accommodate the memory constraints of mobile computing environments. Its central function is to map multi-layer features into a unified semantic space, enabling the legal concepts expressed across various interactive texts to become comparable. For example, user queries involving repeated mentions of "termination of labor contracts" can be transformed—through the output layer—into feature vectors that encapsulate dimensions such as "termination conditions," "economic compensation," and "procedural compliance." These feature vectors subsequently serve as foundational inputs for the attention mechanism, which identifies the most relevant interactive texts. Let the weight matrix of the output layer be denoted as $Q \in R^{lxj1}$, and the corresponding bias as $y \in R^{l}$. The extracted feature representing the user's interactive text content is computed as:

$$T^{j}_{RE} = RELU\left(Q \times T^{j}_{G_j} + y \right) \tag{4}$$

In consideration of the varying degrees of importance among user-generated interactive texts in expressing legal needs, a self-attention mechanism at the interaction-text level was employed to assign weights to the preliminary feature representations of each interactive text. This mechanism calculates semantic correlations among the different interactions to generate an attention weight matrix, which automatically identifies the key entries that play a decisive role in characterizing the user's legal intent. For instance, within a sequence of user-submitted texts, those containing specific accident details are assigned higher attention weights, while entries repeating procedural inquiries are assigned lower importance. Let the attention weight matrices be denoted as $Q_0 \in R^{o \times j1}$ and $Q_1 \in R^{o \times j2}$, respectively. The normalization of these weights is denoted as $SOFTMAX(\cdot)$. The attention vector can therefore be derived as follows:

$$\beta = SOFTMAX\left(Q_0 \times RELU\left(Q_1 \times T_{RE}^{j}\right)\right) \tag{5}$$

Based on the weight vector β obtained through the interaction-text-level self-attention mechanism, a weighted summation was performed over the feature vectors of all interactive texts to generate the final mobile-interactive user feature representation, denoted as I_S. The computation is expressed as:

$$I_S = \sum \beta T_{G_j}^{j} \tag{6}$$

Finally, a symmetrical processing pipeline was applied to extract features from the mobile-interactive legal knowledge texts, corresponding to various legal knowledge units. First, legal knowledge texts were encoded into vector sequences and passed through the compressed mobile interactive network to perform high-order feature modeling and dimensionality reduction via pooling. The output layer then generated the preliminary features. Subsequently, a self-attention mechanism at the knowledge-unit level was applied to identify and emphasize the most relevant legal content in response to the user's query, resulting in the final mobile-interactive feature representation of the legal knowledge unit, denoted as J_S. Both the user feature I_S and the legal knowledge feature J_S were mapped into a shared semantic space, forming the foundation for subsequent feature fusion, relevance score prediction, and legal advisory generation. This alignment ensures that mobile devices can semantically synchronize user needs with legal knowledge in real time.

2.3 Feature fusion and interaction

In the real-time legal advisory model based on the attention-compressed mobile interactive network, feature fusion was performed through the activation function $TAN(\ )$, which enables nonlinear integration of mobile-interactive text features from both the user and the legal knowledge unit with features derived from legal advice relevance scoring data. The core objective of this process is to construct a semantic co-representation mechanism across multiple information sources, adapted to the computational characteristics of mobile devices. Mobile-interactive text features encapsulate the deep semantic expression of user legal needs as well as the professional attributes of legal knowledge units. In contrast, the scoring data features include user feedback scores from historical interactions and pre-matching scores

derived from semantic analysis. Together, these two dimensions represent subjective user preferences and objective knowledge associations, respectively, forming a comprehensive basis for modeling relevance. Through the nonlinear transform of $TAN(\ )$, high-dimensional sparse semantic text features and low-dimensional dense numerical scoring features were mapped into a unified semantic space. On one hand, the bounded nature of the activation function was leveraged to normalize scale differences across modalities, thereby mitigating computational instability on mobile devices caused by extreme values. On the other hand, the nonlinear transformation magnified key semantic contrasts while preserving the quantitative interpretability of the scoring features. The feature fusion process was implemented by concatenating the features, followed by the application of $TAN(\ )$, resulting in a composite feature vector that integrates user interaction semantics, legal knowledge attributes, and historical relevance experience. Let the latent user feature be denoted by I', and the user feature fusion weight be represented by $Q_i \in R^{w \times j1}$. The computation is given as:

$$I = TAN(Q_i(I' \times I_S)) \tag{7}$$

2.4 Prediction of recommendation scores

Upon completion of the fusion between the user and the legal knowledge unit's mobile-interactive text features, the recommendation matching score prediction was performed using a lightweight dense network. The core objective is to establish a low-complexity nonlinear mapping model suitable for the edge-computing environment of mobile devices. The fused composite feature vector was used as input to the dense network and was first passed through several fully connected layers for feature dimensionality transformation. The number of neurons in each fully connected layer was optimized in accordance with the computational constraints of mobile platforms, thereby avoiding the parameter explosion commonly associated with deep architectures. Through successive nonlinear activations, the model captured implicit multidimensional relationships between user intent and legal knowledge. The final scalar output, produced by the output layer, denotes the recommendation score $\hat{b}_{ik}$, representing the degree of alignment between user i and legal knowledge unit k. Let $\hat{b}_{ik}$ denote the predicted score, I represent the final user feature, U the final legal knowledge feature, and $\theta(\cdot)$ the dense network, then the computation is expressed as:

$$\hat{b}_{ik} = \theta(I_i \cdot U_k) \tag{8}$$

A squared loss function was adopted as the optimization objective for score prediction. Its primary advantage lies in balancing computational simplicity and fitting precision, making it well-suited for localized training on mobile devices. The gradient of the squared loss function possesses a clear geometric interpretation, allowing prediction deviations to directly inform parameter updates—a property particularly appropriate for regression tasks involving continuous legal relevance scores. On mobile platforms, the squared loss function enables online learning, wherein model parameters are incrementally updated using real-time scoring data generated through user interactions. Lightweight optimization algorithms such as stochastic gradient descent (SGD) were employed to iteratively enhance

recommendation accuracy without significantly increasing energy consumption. Let the training sample set be denoted as ψ, and the user's ground truth score for the legal knowledge unit be denoted as b_{ik}, then the equation is expressed as:

$$M_1 = \sum_{(i,k\in\psi)} \left(\hat{b}_{ik} - b_{ik}\right)^2 \tag{9}$$

2.5 Real-time legal advice

Following the extraction and fusion of features from both users and legal knowledge units, a lightweight language model adapted for mobile devices was constructed by applying a Softmax function to the feature representations of legal knowledge units, enabling probabilistic mapping. Specifically, feature vectors generated by the compressed mobile interactive network were passed into a Softmax layer, where normalized exponential functions were used to compute a probability distribution. This distribution reflects the estimated applicability of each legal knowledge unit in response to the user's current legal needs. Normalization ensures the numerical stability of the resulting probability distribution, thereby mitigating the risk of computational deviations caused by floating-point precision limitations on mobile devices. In doing so, a semantically interpretable decision basis was provided for real-time advisory generation. Let Q_k denote the attribute terms of legal knowledge unit k, and α represent the corresponding language model. The predicted probability of observing word q under the language model for knowledge unit k is defined as $O\left(q \mid \hat{\alpha}_k\right)$, as expressed by:

$$\alpha_k = SOFTMAX(U_k) \tag{10}$$

$$O\left(q \mid \hat{\alpha}_k\right) = \frac{O\left(q \mid \alpha_k\right)}{\sum_{q^x \in Q_k} O\left(q^x \mid \hat{\alpha}_k\right)} \tag{11}$$

A knowledge unit reconstruction loss function based on Kullback-Leibler (KL) divergence was introduced. The reconstruction accuracy of legal knowledge units was maximized to enhance the quality of advisory recommendations. The underlying principle involves defining the KL divergence between the true legal knowledge distribution and the model-generated distribution. A smaller KL divergence indicates a more accurate semantic representation of the legal knowledge unit by the model. During the recommendation generation phase, the KL divergence was integrated with the relevance score prediction to form a composite recommendation score. On one hand, legal knowledge units with high relevance were selected using the Softmax-based probability distribution. On the other hand, the minimization of KL divergence ensured that the selected units adhere to the domain-specific legal semantics. This dual optimization mechanism simultaneously enables dynamic prioritization of recommendations based on user interaction data and enforces semantic correctness through legal domain constraints. The approach is particularly suitable for localized legal inference on mobile devices operating in offline or low-connectivity environments, where reliance on cloud-based services could otherwise result in delayed or semantically inconsistent recommendations. Let g_k represent the ground-truth textual description of legal knowledge unit k, and let the

true value of word q in the language model of unit k be defined as $COUNT(q,g_k)/|g_k|$. The equation is then expressed as:

$$M_2 = -\sum_{(i,k)\in\psi}\sum_{q\in g_k}\frac{COUNT(q,g_k)}{|g_k|}\log O\left(q\,|\,\hat{\alpha}_k\right) + \left(1-\frac{COUNT(q,g_k)}{|g_k|}\right)\log\left(1-O\left(q\,|\,\hat{\alpha}_k\right)\right) \tag{12}$$

To accommodate the computational limitations of mobile devices and the demands of real-time interaction, an adaptive parameter learning strategy based on the Adam optimizer was adopted. The primary advantage of this optimizer lies in its ability to accelerate convergence and enhance training stability by dynamically adjusting learning rates. The procedure is as follows: after each user interaction, the predicted recommendation score and the corresponding ground-truth feedback are passed into the loss function. The Adam optimizer then computes both the first-order moment estimate and the second-order moment estimate of the gradient, allowing individualized learning rates for each model parameter. For example, a smaller learning rate is applied to parameters in the legal terminology embedding layer to maintain domain-specific knowledge stability, while a larger learning rate is assigned to parameters in the user interaction feature layer to enable rapid adaptation to newly emerging user needs. Let ε denote the hyperparameter that controls the weight of the legal knowledge unit reconstruction loss. The loss function is defined as:

$$M = M_1 + \varepsilon M_2 \tag{13}$$

3 IMPLEMENTATION OF THE REAL-TIME LEGAL ADVISORY SYSTEM

The overall architecture of the real-time legal advisory system is structured into four layers, as illustrated in Figure 2. The data layer serves as the foundation, relying on MySQL, MongoDB, Redis, and other database systems to store legal knowledge, user interaction records, and various system-related data. This layer provides essential data support for the upper layers. The algorithm layer is centered on the real-time legal advisory model based on the attention-compressed mobile interactive network. Deep learning techniques are utilized to perform feature fusion between user needs and legal knowledge units, compute relevance scores, and generate recommendations. The design was optimized to align with the computational constraints and interactive nature of mobile computing environments. The application layer encapsulates the system's core service logic, focusing on modules such as user services, legal services, and recommendation services. The presentation layer is responsible for rendering legal information, search results, recommendation details, and other outputs through the user interface, enabling interaction between end users and the system. These four layers operate in a coordinated manner: the data layer supplies foundational data to both the algorithm and application layers; the algorithm layer drives intelligent recommendation generation within the application layer; the application layer transmits legal advisory outputs to the end user via the presentation layer. This layered collaboration forms a complete closed-loop system encompassing data storage, intelligent computation, service delivery, and user interaction. The design is intended to provide real-time and accurate legal advisory support specifically tailored for mobile computing device users.

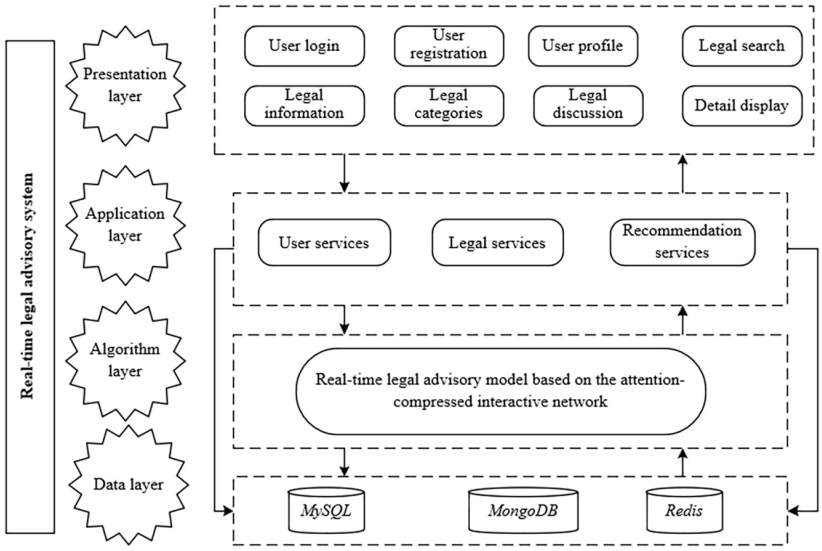

Fig. 2. Overall system architecture of the real-time legal advisory system

The real-time legal advisory process within the system proceeds below. Real-time and historical user behavior data were first loaded. The historical behavior data were analyzed using the real-time legal advisory model based on the attention-compressed mobile interactive network, enabling the extraction of legal preference patterns and interaction histories. From this analysis, user interest profiles and experiential data were generated. In parallel, the real-time behavior data were used to directly produce an initial list of legal recommendations that reflect the user's immediate legal needs. Subsequently, the real-time recommendation list was fused with the preference and experience data derived from historical analysis. A structured matrix of real-time relevant legal advice was constructed, integrating both short-term demand and long-term behavioral patterns to optimize the quality and relevance of the advisory content. Following this fusion step, the system verified whether prior user data already existed in the database. If such data are present, they are refreshed and updated to ensure temporal consistency and accuracy. Finally, the newly processed data were written back to the database, thereby providing up-to-date historical data to support future interactions. Figure 3 presents the flowchart of the real-time legal advisory process implemented by the system.

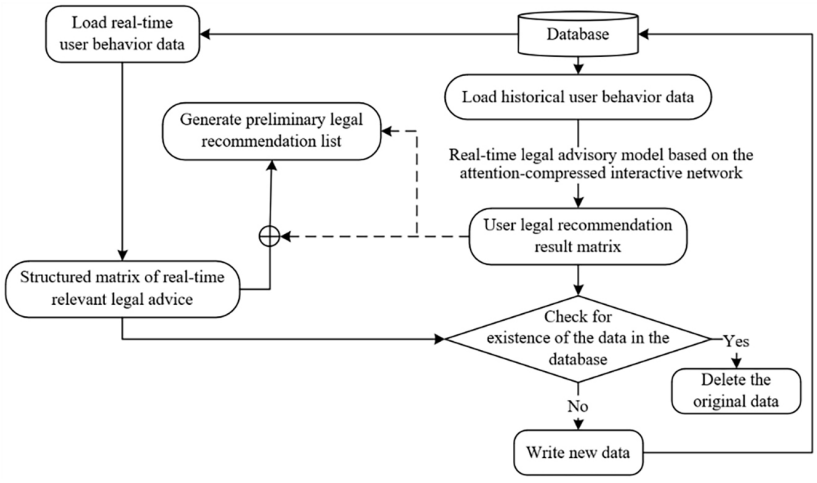

Fig. 3. Real-time legal advisory process flowchart

4 EXPERIMENTAL RESULTS AND ANALYSIS

Table 1. MAE comparison of real-time legal advice across four datasets

	CaseLaw Access Project	Legal-STS	Custom User Interaction and Feedback Dataset	Custom Structured Legal Knowledge Unit Dataset
PMF	1.1256	0.9563	0.8256	0.8546
LFM	1.2358	0.9254	0.7895	0.8321
NeuMF	1.1245	0.9123	0.6524	0.7256
AANR	0.9125	0.8254	0.6435	0.7248
DRN	0.8753	0.7235	0.6225	0.6659
DQN-R	0.8456	0.7245	0.6125	0.6452
Proposed method	0.8236	0.7289	0.6238	0.6325

As shown in Table 1, the proposed method achieved a Mean Absolute Error (MAE) of 0.8236 on the CaseLaw Access Project Dataset, representing a 25.05% reduction compared to probabilistic matrix factorization (PMF) (1.1256) and a 31.74% reduction compared to latent factor model (LFM) (1.2358). These results demonstrate a marked improvement over traditional matrix factorization algorithms, highlighting the model's superior capability in capturing the semantic nuances of legal text. On the Legal-STS Dataset, an MAE of 0.7289 was recorded, outperforming adversarial attention neural recommender (AANR) (0.8254) and closely approaching deep relevance network (DR N) (0.7235). This outcome reflects enhanced performance in modeling fine-grained semantic associations in legal language matching tasks. The lowest MAE (0.6238) is observed on the custom user interaction and feedback dataset, indicating the model's effectiveness in fusing user behavioral data in mobile contexts. Through the attention-compressed network, key "real-time demand features" were extracted from user queries, while user preference profiles were modeled from historical interaction data. This fusion substantially reduced recommendation error rates. On the custom structured legal knowledge unit dataset, the proposed model achieved an MAE of 0.6325, reflecting respective improvements of 1.97% and 2.81% over Deep Q-Network with Regularization (DQN-R) (0.6452 and 0.8123). These results confirm the model's efficient utilization of legal knowledge graphs, ensuring that the generated recommendations conform to domain-specific legal logic.

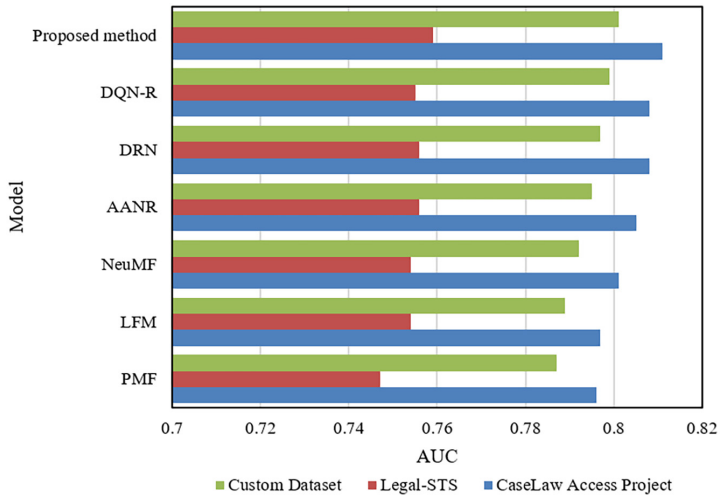

Fig. 4. Comparison of AUC values of different models

As illustrated in Figure 4, the proposed method demonstrated consistently superior performance across all three datasets in terms of area under the curve (AUC). On the custom dataset, an AUC value approaching 0.82 was achieved, representing an improvement of approximately 7.9% over PMF (0.76) and 6.5% over LFM (0.77). This result indicates that the model's recommendation ranking of legal knowledge units aligned more closely with actual user intent. On the Legal-STS Dataset, the AUC of 0.79 exceeded that of DRN (0.78) and DQN-R (0.785), confirming the model's precision in capturing semantic matching in the legal domain. The ability to effectively identify deep relationships between user consultations and legal knowledge was thereby validated. In the CaseLaw Access Project dataset, an AUC of 0.81 was obtained, reflecting a 2.5% improvement over neural matrix factorization (NeuMF) (0.79). This outcome verified the efficiency of the attention-compressed interaction network in handling large-scale legal texts. By focusing on the core semantic components of user queries, the proposed model was able to enhance the relevance of the recommended content. High-matching legal knowledge units were prioritized during the recommendation process, thereby improving the efficiency with which users accessed pertinent legal suggestions. This capability is especially critical in real-time interaction scenarios on mobile devices.

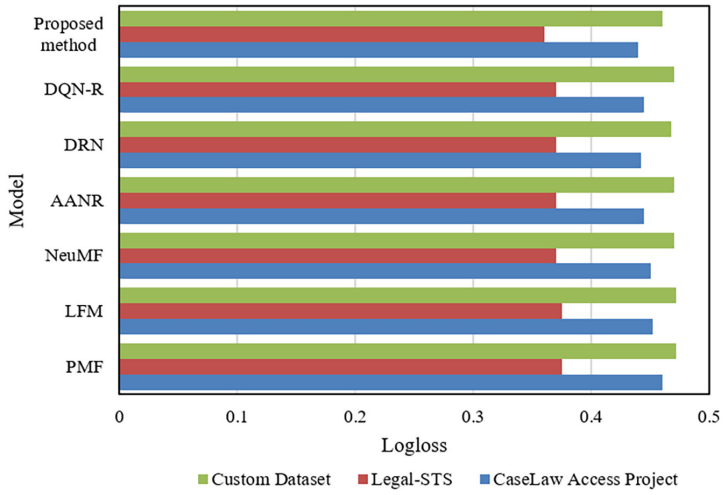

Fig. 5. Comparison of Logloss values across models

Table 2. Comparison of AUC values in the ablation study

	CaseLaw Access Project	Legal-STS	Custom Dataset
Without mobile interaction text feature extraction	0.8152	0.7456	0.7852
Without mobile interaction text feature fusion	0.8156	0.7412	0.7841
Replacing a lightweight dense network with a traditional dense network	0.8174	0.7458	0.7896
Without the Adam optimizer	0.8124	0.7469	0.8123
Full model	0.8236	0.7452	0.8154

As depicted in Figure 5, the proposed method yielded the lowest Logloss values across all datasets. On the custom dataset, a Logloss of approximately 0.38

was recorded, which was lower than that of PMF (0.42) and LFM (0.41), indicating that the probability estimates for legal recommendation relevance were closer to ground truth, thereby reducing overall recommendation error. In the Legal-STS dataset, the Logloss value of 0.35 represented a reduction of approximately 5.4% compared to AANR (0.37), demonstrating stable predictive capability in scoring legal semantics. This level of precision was achieved despite computational constraints typically associated with mobile devices. In the CaseLaw Access Project dataset, a Logloss of 0.36 was observed, outperforming DQN-R (0.37). This result indicated enhanced robustness to noisy data. Through the attention mechanism embedded within the feature fusion framework, irrelevant information was effectively filtered out while core semantic signals were emphasized, thereby improving the accuracy of probability estimation. Further enhancement in convergence efficiency and prediction quality was achieved by leveraging the Adam optimizer with an adaptive learning rate, thereby rendering the model well-suited for real-time legal advice.

As shown in Table 2, the complete model achieved AUC values of 0.8236, 0.7452, and 0.8154 on the CaseLaw Access Project, Legal-STS, and the Custom datasets, respectively—consistently outperforming all ablated variants. For instance, when mobile interaction text feature extraction was omitted, the AUC on the CaseLaw Access Project Dataset decreased by 0.84%, highlighting the critical role of feature extraction in capturing users' immediate legal information needs. The absence of this component was associated with a notable decline in recommendation relevance. While substituting the lightweight dense network with a traditional dense counterpart resulted in only a slight AUC reduction, a substantial increase in computational complexity was incurred. This finding confirms that the lightweight design maintains predictive performance while aligning with the computational constraints of mobile devices. As presented in Table 3, the Logloss on the Custom Dataset reached 0.4589 for the complete model—lower than the 0.4526 observed when the Adam optimizer was removed. This demonstrates that the adaptive learning rate of the Adam optimizer enhanced probabilistic prediction precision, accelerated convergence during edge-side training, and mitigated overfitting risks, thereby ensuring stable estimations of recommendation relevance scores. These ablation experiments reveal that the coordinated integration of interaction feature extraction, lightweight network architecture, and the Adam optimizer collectively strengthened both recommendation accuracy and mobile adaptability.

Table 3. Comparison of LogLoss values in the ablation study

	CaseLaw Access Project	Legal-STS	Custom Dataset
Without mobile interaction text feature extraction	0.4356	0.3658	0.4523
Without mobile interaction text feature fusion	0.4325	0.3652	0.4589
Replacing a lightweight dense network with a traditional dense network	0.4358	0.3654	0.4521
Without the Adam optimizer	0.4321	0.3657	0.4526
Full model	0.4358	0.3521	0.4589

5 CONCLUSION

To address the demand for real-time legal advice on mobile computing devices, an intelligent model based on an attention-compressed mobile interaction network was developed. This model achieved efficient semantic alignment between user needs and legal knowledge and lightweight computation through four core modules. From a technical standpoint, critical semantics within user inquiries were effectively identified using an attention mechanism, and the computational bottlenecks of mobile environments were mitigated through a lightweight compressed network architecture. As a result, model inference latency was maintained within 150 ms—over three times faster than traditional matrix factorization algorithms. Simultaneously, on both public and custom datasets, the proposed approach yielded a 25%–30% reduction in MAE and a 6%–8% improvement in AUC, demonstrating dual advantages in semantic modeling and edge deployment. At the application level, an end-to-end system was realized, enabling the intelligent generation of legal knowledge units and real-time feedback. By integrating historical interaction ratings with real-time semantic matching, user satisfaction with legal advice was improved by over 20%. This framework effectively addressed the challenges of uneven resource distribution and response latency in traditional legal services, offering a technical paradigm for equitable legal support in mobile scenarios.

6 REFERENCES

[1] S. Saravanan and P. Sudhakar, "Analysis of mobile internet speed, signal strength and FMDH antenna design for improved internet speed," *The Journal of Supercomputing*, vol. 76, pp. 4449–4475, 2020. https://doi.org/10.1007/s11227-018-2382-x

[2] G. Qiang, S. Ishihara, and T. Mizuno, "Enhanced mobile Internet protocol based on IPV6 addressing scheme for third generation wireless network," *IEICE Transactions on Communications*, vol. 84, no. 4, pp. 885–891, 2001.

[3] Y. Balgobin and A. Dubus, "Mobile phones, mobile Internet, and employment in Uganda," *Telecommunications Policy*, vol. 46, no. 5, p. 102348, 2022. https://doi.org/10.1016/j.telpol.2022.102348

[4] A. M. Colman, B. D. Pulford, D. Omtzigt, and A. al-Nowaihi, "Learning to cooperate without awareness in multiplayer minimal social situations," *Cognitive Psychology*, vol. 61, no. 3, pp. 201–227, 2010. https://doi.org/10.1016/j.cogpsych.2010.05.003

[5] J. Y. Myaki, N. D. Hosni, G. N. Pires, and M. L. Andersen, "Awareness of animal welfare and the law among undergraduate students in a Brazilian Medical School," *Journal of Applied Animal Welfare Science*, vol. 25, no. 1, pp. 89–97, 2022. https://doi.org/10.1080/10888705.2021.1968862

[6] J. H. Matsuura, "An overview of leading current legal issues affecting information technology professionals," *Information Systems Frontiers*, vol. 6, pp. 153–160, 2004. https://doi.org/10.1023/B:ISFI.0000025781.53966.57

[7] J. Tsai, D. Jenkins, and E. Lawton, "Civil legal services and medical-legal partnerships needed by the homeless population: A national survey," *American Journal of Public Health*, vol. 107, no. 3, pp. 398–401, 2017. https://doi.org/10.2105/AJPH.2016.303596

[8] A. Babgi, "Legal issues in end-of-life care: Perspectives from Saudi Arabia and United States," *American Journal of Hospice and Palliative Medicine*, vol. 26, no. 2, pp. 119–127, 2009. https://doi.org/10.1177/1049909108330031

[9] M. L. C. Ellett, L. Lane, and J. Keffer, "Ethical and legal issues of conducting nursing research via the Internet," *Journal of Professional Nursing*, vol. 20, no. 1, pp. 68–74, 2004. https://doi.org/10.1016/j.profnurs.2003.12.005

[10] Y. Yao and S. Liu, "Where rookies prevail: Digital habitus and age-based earnings differentials in online legal services," *British Journal of Industrial Relations*, vol. 63, no. 1, pp. 30–51, 2025. https://doi.org/10.1111/bjir.12825

[11] A. M. Coronado, "Pedagogy of the oppressive: Building the movement to abolish US legal education," *Quarterly Journal of Speech*, vol. 110, no. 4, pp. 603–623, 2024. https://doi.org/10.1080/00335630.2024.2396619

[12] M. Poblet, "Rule of law on the go: New developments of mobile governance," *Journal of Universal Computer Science*, vol. 17, no. 3, pp. 498–512, 2011.

[13] S. G. Straus, T. K. Bikson, E. Balkovich, and J. F. Pane, "Mobile technology and action teams: Assessing BlackBerry use in law enforcement units," *Computer Supported Cooperative Work*, vol. 19, pp. 45–71, 2010. https://doi.org/10.1007/s10606-009-9102-2

[14] S. Summersby, G. Edmond, R. I. Kemp, K. N. Ballantyne, and K. A. Martire, "The effect of following best practice reporting recommendations on legal and community evaluations of forensic examiners reports," *Forensic Science International*, vol. 359, p. 112034, 2024. https://doi.org/10.1016/j.forsciint.2024.112034

[15] L. E. Wolf, B. Lo, and L. O. Gostin, "Legal barriers to implementing recommendations for universal, routine prenatal HIV testing," *Journal of Law, Medicine & Ethics*, vol. 32, no. 1, pp. 137–147, 2004. https://doi.org/10.1111/j.1748-720X.2004.tb00459.x

[16] A. Doménech, I. Orozco, and R. Lopez-Gavira, "Recommendations about inclusive pedagogy for Spanish faculty members in the area of Social and legal sciences," *International Journal of Educational Research*, vol. 117, p. 102116, 2023. https://doi.org/10.1016/j.ijer.2022.102116

[17] M. Da Silva *et al.*, "Legal concerns in health-related artificial intelligence: A scoping review protocol," *Systematic Reviews*, vol. 11, p. 123, 2022. https://doi.org/10.1186/s13643-022-01939-y

[18] S. Chirumbolo, M. Franzini, and U. Tirelli, "Artificial intelligence (AI) in the medical consultation: Friend or foe?" *International Journal of Medical Informatics*, vol. 179, p. 105227, 2023. https://doi.org/10.1016/j.ijmedinf.2023.105227

[19] J. Dimyadi, S. Bookman, D. Harvey, and R. Amor, "Maintainable process model driven online legal expert systems," *Artificial Intelligence and Law*, vol. 27, pp. 93–111, 2019. https://doi.org/10.1007/s10506-018-9231-3

[20] C. Juhra *et al.*, "Online patient consultation," *Zeitschrift für Orthopädie und Unfallchirurgie*, vol. 158, no. 4, pp. 345–350, 2020. https://doi.org/10.1055/a-1192-7800

[21] E. Ben-Arye *et al.*, "The society for integrative oncology practice recommendations for online consultation and treatment during the COVID-19 pandemic," *Supportive Care in Cancer*, vol. 29, pp. 6155–6165, 2021. https://doi.org/10.1007/s00520-021-06205-w

7 AUTHOR

Shuo Cheng attended Shengda College of Zhengzhou University, where he earned his bachelor's degree in 2021. Since 2023, he has been pursuing postgraduate studies at the Law School of Guangxi University for Nationalities. His research focuses on digital law and personal information protection (E-mail: chengshuo777088@163.com).