Deep Learning based Dialogue System for Legal Consultancy in Smart Law

Xukang Wang[1]
Sage IT Consulting Group
Shanghai, China
xukangwang@sageitgroup.com

Ying Cheng Wu[2*]
School of Law
University of Washington
Seattle, USA
wyc9@uw.edu

Xuhesheng Chen[3]
The University of North Carolina at Chapel Hill
Chapel Hill, USA
xuhesheng.chen@alumni.unc.edu

Hongpeng Fu[4]
Tandon School of Engineering
Northeastern University
Seattle, USA
fuhp@pku.edu.cn

Jiaqi Tan[5]
School of Computing Science
Simon Fraser University
Burnaby, Canada
jiaqit@sfu.ca

Mengjie Zhou[6]
Department of Computer Science
University of Bristol
Bristol, UK
mengjiezhou2018@outlook.com

***Abstract*—The integration of artificial intelligence (AI) into the legal domain has emerged as a prominent area of research. Notably, AI's application within legal consulting has the potential to markedly enhance the precision and efficiency of counsel. Nevertheless, the intricate nature of legal statutes and the unique nuances of individual cases pose considerable challenges in developing efficacious AI-driven legal consultation systems. This study introduces a deep learning-powered dialogue system designed to comprehend user inquiries and deliver precise legal advice. Empirical evidence attests to the system's superior performance in legal advisory functions, thereby substantiating its role in augmenting the effectiveness and accuracy of legal consultation services.**

Keywords—Artificial Intelligence, Legal Consultation, Deep Learning, Dialogue System, Smart Laws

I. INTRODUCTION

Legal consultancy services are an indispensable part worldwide [1]. In civil cases such as property disputes, divorce lawsuits, inheritance issues, or criminal cases such as suspected crime defense, individuals typically require professional legal advice to guide their actions. This demand exists not only in everyday life but also holds a central position in the business environment [2]. Legal consultation can help enterprises operate in compliance with various industry regulations and avoid business risks. However, since traditional legal consultancy is centered on manual services by lawyers, its efficiency and speed are limited by the workload and time of lawyers. Moreover, the differences in professional capabilities and experience of lawyers can lead to some fluctuations in the accuracy of consultation. With the rapid development of socio-economy and the increasing complexity of the legal environment, how to improve the efficiency and accuracy of legal consultation through technological means, especially AI technology, has become an important societal demand [3].

Despite AI having achieved significant breakthroughs in fields such as language translation, medical diagnosis, financial analysis, applying it to the field of legal consultation still faces major challenges [4]. Firstly, law is a professional discipline covering various complex fields, including criminal law, civil law, business law, administrative law, among others. Each branch has a large number of legal provisions, regulations, jurisprudences, and precedents, as well as many legal terms and legal statements [5]. These all pose extremely high requirements for the knowledge understanding and processing capabilities of the AI system [6]. For instance, the AI system needs to understand the meaning of legal provisions, parse legal terms, and even understand the correlation and logic between legal provisions. Additionally, the process of legal consultation often involves a large number of vague and specific case issues, such as “Is it legal for me to do this?” and “How should I operate to maximize the protection of my rights?” [7] This requires the AI system to have strong case analysis and reasoning capabilities, and even requires understanding and simulating the situation. These pose significant challenges to the design and training of the AI system.

In response to the challenges of legal consultation, this study designed a dialogue system based on deep learning to handle legal consultation tasks. This research builds on and expands the work of previous studies. For example, we are indebted to the researchers who have pioneered AI applications in legal consultancy services. Firstly, the system uses natural language processing (NLP) technology to understand and parse user consultation content. This part not only includes basic grammar parsing and semantic understanding, but also understanding the implied meaning in the consultation content, and even the possible legal issues involved. Secondly, our system conducts reasoning in conjunction with the legal knowledge base to provide legal consultancy. The legal knowledge base is manually collected and labeled by our team, covering various legal fields in the

United States, including legal provisions, regulations, jurisprudence, and precedents. Based on this legal knowledge, our system can perform logical reasoning and simulated judgments to provide legal consultancy. However, legal consultation often involves many vague and specific case issues, which requires the AI system to have strong case analysis and reasoning capabilities, and even to understand and simulate the situation. Therefore, this study has also designed a special reasoning algorithm that can provide the best legal consultation for users based on existing legal knowledge and similar cases. This reasoning algorithm combines case reasoning and rule reasoning, capable of dealing with complex, vague, and specific case-related legal problems. In addition, our system supports continuous learning and improvement. As usage deepens, the system can continuously learn and adjust, obtain information from user feedback and evaluations, and improve the accuracy and efficiency of consultation through training and optimization. This is not achievable in traditional legal consultation services and is a major advantage of our system.

The contributions of this research mainly lie in the following aspects: Firstly, this study designed and implemented an AI legal consultation system with highly specialized capabilities. During the design process, the researchers referred to a large amount of legal theory and practical experience, as well as the latest research results in the field of AI, thereby ensuring the theoretical and practical nature of the system [8]. This is also a major innovation of this research. Secondly, the system can not only provide efficient legal consultation services, but also handle complex, vague, specific case-related legal problems, which are difficult to achieve in traditional legal consultation services. This shows that AI technology has a great application prospect and potential in the legal field. In addition, this research provides valuable practical experience for AI applications in the legal field, providing reference and inspiration for future research and development. This research also shows that AI technology can be combined with traditional professional fields to create new service models and values. Lastly, this research also has great social value. By improving the efficiency and accuracy of legal consulting, our AI system can provide legal help to more people, democratizing access to legal advice.

II. System Design

This section will delve into the design of our AI legal consulting system. The overall architecture of the system consists of three main parts: the user interaction interface, the knowledge understanding and reasoning engine, and the learning and optimization module.

A. User Interaction Interface

The user interaction interface is the platform for users to interact with the system. This study designed a user-friendly interactive interface, where users can input their legal questions and see the legal consultation results generated by the system. The user interaction interface adopts a Q&A design. Users can query by typing in questions, and the system will provide the corresponding answers. To optimize the user experience, we considered usability and comprehensibility in our interface design. We offer some preset question templates on the interface to help users better describe their legal issues. Simultaneously, the system will automatically suggest related questions or advice based on user input, guiding users to provide more background or details to assist the system in understanding and answering the questions more accurately.

In the interaction design, this study adopted a natural language dialogue mode so that users can pose their legal issues naturally and routinely. We support not only single-question queries but also dialogue-style interactions. For example, users can initially pose a more general question, and based on the system's response, they can pose more specific questions if needed. This type of interaction can simulate real consulting scenarios, helping users get more in-depth answers.

To better describe their legal issues, we provide some preset question templates on the user interaction interface. These templates cover a variety of common legal issues, and users can directly select them and add or modify specific details based on the template. This design simplifies user input and helps them describe their issues more accurately.

In terms of output, our system will generate corresponding legal consultation results according to the type of user questions. These results are displayed in an easy-to-understand manner, including text descriptions, lists, charts, etc. When necessary, our system will also provide corresponding legal provisions or case law for reference. In addition, our system supports result download and sharing. Users can save the query results in various formats (such as PDF, Word, etc.) or directly share them on social media. In this way, users can conveniently access or share their query results when needed. We will detail the system's knowledge understanding and reasoning engine in the following section.

Overall, the user interaction interface is the front end of the system. Users can interact with our AI legal consulting system through this interface. Our design goal is to make all users, whether they have a legal background or not, able to easily use our system. Hence, we put a lot of effort into designing the user interaction interface to ensure its usability and comprehensibility.

B. Knowledge Understanding and Reasoning Engine

The knowledge understanding and reasoning engine is at the core of the AI legal consulting system. This part includes two main modules: the natural language processing module and the legal knowledge base. Together, these modules realize functions such as user question understanding, legal knowledge searching, and reasoning.

(1) Natural Language Processing Module

The natural language processing module is key to understanding user questions [9]. User-proposed questions usually involve various language phenomena such as vague semantics, complex grammatical structures, and long-distance dependencies. To accurately understand these questions, this study employs the latest deep learning and

natural language processing technologies. Our natural language processing module mainly includes three steps: text preprocessing, feature extraction, and semantic understanding.

(2) Legal Knowledge Base

The legal knowledge base is the knowledge source for the system to provide legal consultation. Our legal knowledge base includes a large number of legal provisions, regulations, jurisprudence, case laws, and related metadata (such as the effective date, the scope of impact, etc.).

(3) Knowledge Reasoning

After understanding the user questions and finding related legal knowledge, our system needs to carry out knowledge reasoning to generate legal consultation results.

In summary, the knowledge understanding and reasoning engine is the core part of the system. This part consists of two main components: the natural language processing module and the legal knowledge base. The natural language processing module is responsible for handling user-input legal questions, understanding the meaning of the question, analyzing the background of the question, extracting key information, and identifying related legal concepts and entities. This module uses deep learning and natural language processing technologies and can handle various complex language phenomena. The legal knowledge base is the main source of knowledge for the system to provide legal consultation. The legal knowledge base includes a large number of legal provisions, regulations, jurisprudence, case laws, etc. This legal knowledge is stored in a structured form, which is convenient for searching and reasoning. The system will find relevant knowledge from the legal knowledge base based on the analysis results of user questions, and then provide advice or consultation. Based on these two modules, our system can handle user legal questions, understand the meaning of the question, analyze the background of the question, find relevant legal knowledge, and then reason based on this knowledge to generate consultation results.

C. Learning and Optimization Module

To ensure that our AI legal advisory system can self-improve and continuously enhance its performance, we have designed a learning and optimization module. This module is mainly composed of two parts: model training and online optimization.

(1) Model Training

The system needs to be able to understand and deal with a variety of legal issues, so this study uses a supervised learning method to train our system. This study first collects a large number of training samples from historical consulting records, and then uses these samples to train the model. Each training sample contains a user question, one or more relevant legal knowledge, and a consulting result. These samples allow the system to learn how to correctly match user questions with relevant legal knowledge and generate accurate consulting results based on this. To deal with the complexity and diversity of user questions, the model uses deep learning techniques. Specifically, this study adopts the Transformer structure and self-attention mechanism, which can capture long-distance dependencies and handle complex matching patterns.

In the model training process, this study uses a loss function to measure the gap between the model's predicted results and the actual results. To train the model, we use the cross-entropy loss function, the formula of which is as follows:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Where N represents the number of training samples, y^i represents the actual result of the i-th sample, and $\hat{y}_i$ represents the model's prediction of the i-th sample. We use the gradient descent method to minimize the loss function and train the model parameters. In each iteration, we update the parameters based on the gradient of the loss function to improve the model's prediction effect.

(2) Online Optimization

In addition to model training, this study has also designed an online optimization strategy that allows the system to self-improve based on real-time user feedback. This study adopts a strategy called Multi-Armed Bandit, which allows the system to balance between exploring new legal knowledge and exploiting known legal knowledge. In the Multi-Armed Bandit algorithm, each legal knowledge is considered an arm of a bandit. Whenever there is a user question that needs to be answered, the system needs to select an arm, i.e., select legal knowledge to answer the question. The system selects arms based on the historical effects of each arm but also explores arms with unknown effects at a certain probability. To implement this strategy, this study uses the Upper Confidence Bound algorithm (UCB) [10]. The UCB algorithm chooses the arms with the highest upper confidence bound, i.e., those arms with good historical effects and potentially better effects. The formula for the UCB algorithm is as follows:

$$A_t = arg\max_a [Q_t(a) + c\sqrt{\frac{\log(t)}{N_t(a)}}]$$

Where A_t represents the arm selected at time t, $Q_t(a)$ represents the average gain of arm a, $N_t(a)$ denotes the number of times arm a has been selected up to time t, and c is a parameter that controls the balance between exploration and exploitation. Through model training and online optimization, our system can continuously learn from user feedback and continuously improve its consulting effect. In the next section, we will detail the performance of our system in practical applications.

III. ALGORITHM DESIGN

The core of our AI legal consultation system is the design of its deep learning algorithm. This design takes into account the handling of user queries, the integration of relevant legal knowledge, and the generation of final consultation results. In this section, this study provides a detailed description of our algorithm design, including the model selected, how to handle input and output, as well as our training and optimization strategies.

A. Model Design

The core of our system is a deep learning model based on Transformer [11]. Transformer models have demonstrated superior performance in various NLP tasks, particularly in dialogue systems [12]. This is primarily due to its self-attention mechanism that effectively handles long-range dependencies in the sequence. The Transformer model is a sequence-to-sequence model based on the attention mechanism, consisting of an encoder and a decoder [13]. The encoder receives the input sequence and generates a continuous hidden state sequence [14]. The decoder then generates the output sequence based on the encoder's hidden state sequence and its own historical hidden states [14]. In our system, this study made some customizations to the Transformer model. Specifically, we introduced a representation of legal knowledge in the encoder to enhance the model's legal consultation capability. In the decoder, we adopted a beam search strategy to generate multiple candidate consultation results. Figure 1 represents the transformer.

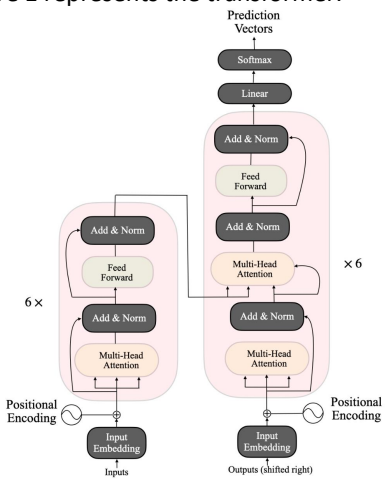

Figure 1. The framework of our proposed model

The Transformer architecture, pivotal to this study, comprises an encoder and a decoder. Specifically, the encoder transforms an input sequence ($x_1, x_2, \ldots, x_n$) into a series of continuous hidden representations ($z_1, z_2, \ldots, z_n$). Building upon these representations, the decoder generates the corresponding output sequence ($y_1, y_2, \ldots, y_n$). Our implementation utilizes six homogeneous layers for both the encoder and the decoder. Each encoder layer is composed of a multi-head self-attention mechanism and a position-wise feed-forward network, augmented with residual connections. Similarly, a decoder layer mirrors the encoder structure with the addition of an encoder-decoder attention mechanism, which aligns the decoder's current state with the entire output from the encoder. To ensure causal decoding, the decoder's input is strategically masked to prevent the contingent prediction on future tokens. Positional encodings are incorporated into both the encoder and decoder inputs to imbue the model with a sense of sequential order, which is essential for processing sequence data without recurrence.

For an input sentence $x = (x_1, x_2, \ldots, x_n)$, each token X_i corresponds to three vectors: query, key, and value. The self-attention computes the attention weight for every token x_i against all other tokens in x_i by multiplying the query of x_i with the keys of all the remaining tokens one by one. For parallel computing, the query, key, and value vectors of all tokens are combined into three matrices: Query(Q), Key(K), and Value(V). The self-attention of an input sentence X is computed by the following equation:

Within the Transformer model, each token x_i of an input sentence $x = (x_1, x_2, \ldots, x_n)$ is mapped to a trio of vectors—query (Q), key (K), and value (V). The self-attention mechanism calculates the attention score of each token X_i with respect to every other token in the sentence by taking the dot product of x_i's query vector with the key vectors of all tokens. To facilitate parallel computation, the query, key, and value vectors for all tokens are aggregated into their respective matrices Query(Q), Key(K), and Value(V). The attention function is applied across these matrices as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Here, the *softmax* function is applied to the scaled dot product of the query and key transposed, normalized by the square root of the dimensionality of the key vectors $\sqrt{d_k}$, ensuring that the attention weights are distributed in a stable manner. This result is then multiplied by the value matrix V to yield the weighted sum output of the self-attention for the input sentence.

To jointly consider the information from different subspaces of embedding, query, key, and value vectors are mapped into h vectors of identical shapes by using different linear transformations, where h denotes the number of heads. Attention is computed on each of these vectors in parallel, and the results are concatenated and further projected. The multi-head attention can be described as:

In the multi-head attention mechanism, the model simultaneously projects the query, key, and value vectors into h distinct representation subspaces using separate linear transformations. This approach allows the model to capture information from different perspectives at different positions. The attention scores are computed in parallel across these h transformed sets of vectors, known as 'heads'. The resulting attention outputs are then concatenated and subjected to a final linear transformation. The multi-head attention mechanism can be formally represented as follows:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^0$$

Where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, W_i^Q, W_i^K, and W_i^V are the parameter matrices for the i-th head, W^0 denotes the output linear transformation matrix.

The Transformer architecture, as proposed, eschews recurrent units, thereby omitting inherent sequence order information. To counter this, positional encodings are amalgamated with input embeddings to impart necessary positional context. The positional encoding adopted in this work utilizes cosine functions, which are formulated as follows:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right)$$

Where *pos* denotes the position of the target token and *i* denotes the dimension, which means that each dimension of the positional matrix uses a different wavelength for encoding.

B. Handling of Input and Output

The inputs to our model include user queries and relevant legal knowledge. User queries are natural language texts. This study uses a pre-trained word embedding model (such as Word2Vec [15]) to convert them into vector representations, which are then fed into our model. Our system also has a legal knowledge base that contains a large number of legal provisions and cases. Our system selects relevant legal knowledge based on the user's question and uses the same word embedding model to convert this legal knowledge into vector representations. The output of our model is the consultation result, which is also a natural language text. Our model uses a generative approach to generate consultation results, i.e., it generates the next word at each step based on the current state and historical consultation results. Specifically, our model generates a probability distribution of words at each step, and samples a word from this distribution as the output for the current step.

C. Training and Optimization

We use a supervised learning approach to train our model. Our training data consists of a large number of user queries, relevant legal knowledge, and consultation results. We use a cross-entropy loss function to measure the difference between the model's predicted results and the actual results. The formula for the cross-entropy loss function is as follows:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Where y_i represents the probability distribution of the actual result, and $\hat{y}_i$ represents the probability distribution of the model's predicted result. We use the Adam optimizer to train our model [16]. The Adam optimizer combines the advantages of the Momentum and RMSProp [17] optimization algorithms and can effectively handle irregularities in parameter updates. The update rules for the Adam optimizer are as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

$$\theta_t = \theta_{t-1} - \alpha \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Where θ_t is the parameter at time *t*, g_t is the gradient at time *t*, m_t and v_t are estimates of the first and second moments of the gradient, α is the learning rate, β_1 and β_2 are the decay coefficients for the moment estimates, and ϵ is a small constant added to prevent division by zero. In addition to model training, this study also designed an online optimization strategy that allows our system to self-improve based on real-time user feedback. This study uses the UCB algorithm to select legal knowledge, thus achieving a balance between exploration and exploitation.

IV. EXPERIMENT

A. Dataset

Our methodological assessment is benchmarked against the MHLCD dataset [18], a recently introduced resource tailored to encapsulate the legal advisory conversational sphere, particularly for women and children subjected to criminal acts (refer to Table 1 for a synopsis). This dataset encompasses dialogues related to legal support for individuals affected by a spectrum of criminal activities, such as domestic abuse, sexual assault, acid attacks, both physical and digital forms of stalking, harassment in the workplace and online, identity fraud, trolling, matrimonial deception, financial deceit, child explicit content, trafficking of women and minors, non-consensual sharing of intimate images, privacy breaches, and various forms of discrimination.

Table 1. Details of the MHLCD dataset

Metrics	Training	Validation	Test
Number of Dialogues	755	100	151
Number of Utterances	20886	2795	4163
Avg. Utterances per Dialogue	27.66	27.95	27.57

B. Evaluation metrics

The assessment of the suggested approach incorporates both automatic and human evaluation metrics. The efficacy of classifiers for counseling tactics, courtesy, and compassion is measured by Weighted Accuracy (W-ACC) and Macro-F1 scores, which are specifically chosen to adjust for disparities in class distribution [18].

This study ascertains the efficiency of the model by gauging its performance based on task accomplishment (in this context, counseling, civility, and compassion) as well as the quality of the responses it generates. The metrics for task achievement include: CoStr, which tallies the utterances articulated with a counseling strategy; Pol, quantifying the frequency of courteous expressions; and Emp, calculating the instances of empathetic utterances produced. To evaluate the response quality, we examine the Perplexity (PPL) score and the response length (R-LEN). CoStr, Pol, and Emp are measured using respective classifiers for counseling, politeness, and empathy. The precision of these classifiers on a test dataset delivers the corresponding metric scores (CoStr, Pol, Emp) for our proposed technique.

Human evaluation is done by recruiting six evaluators with postgraduate qualification and proficiency in similar tasks. To test the robustness of our system, each evaluator is asked to interact with our system 3 times, with a constraint that each time they would have to interact by

using a different set of responses. Then, these 18 human-evaluated dialogues are sent to the experts from government-run institutions for cross-verification in terms of evaluation quality. After experts pass the evaluation process, further 42 dialogues are evaluated. Hence, we end up with total 60 human evaluated dialogues. All six evaluators are asked to rate each dialogue interaction in terms of counseling strategy correctness (Con), politeness (Pol), empathy (Emp), consistency (Const), fluency (Fluen), and non-repetitiveness (N-Rep) on an integer scale of 1-5.

C. Comparison methods

Our comparative analysis encompasses six benchmark models, namely LSTM [19][20], ByteNet [21][22], ConvS2S [23][24], S2S+attention [25][26], and PGN [27][28].

D. Experimental settings

The conducted experiments utilized a single NVIDIA 4090 GPU for training purposes. The word embedding dimension was established at 300, with word2vec employed to generate the initial word vectors. Role embeddings were arbitrarily initiated with a dimensionality of 100. For the model architecture, a 300-dimensional hidden layer size was chosen, and a 6-layer Transformer with eight heads was implemented. We set the dropout rate at 0.8 to mitigate overfitting. Optimization of the objective function was performed using a learning rate of 5e-4, and mini-batch gradient descent was executed with batches of 64. In the decoding phase, the maximum utterance length was restricted to 40 to accommodate potential sub-utterances within the generated sentences.

V. Result and Discussion

A. Evaluation of comparison methods

Figure 2 delineates the distribution of W-ACC, and Macro-F1 for each model under investigation. Evaluation results of all comparison methods are shown in Table 2. It can be observed that the proposed method achieved significantly superior scores in terms of both W-ACC and Macro-F1. Furthermore, the W-ACC of the proposed approach achieves 92.62%. This represents a substantial improvement of approximately 12.5% over the least effective model, LSTM. Such results emphatically affirm the efficacy and essentiality of employing the advanced model for the task of legal dialogue system.

Table 2 Evaluation results of comparison methods

Model	W-ACC(%)	Macro-F1(%)
LSTM	82.36	78.91
ByteNet	86.22	85.32
ConvS2S	88.67	86.17
S2S+attention	88.96	87.38
PGN	89.57	88.69
This study	**92.62**	**91.53**

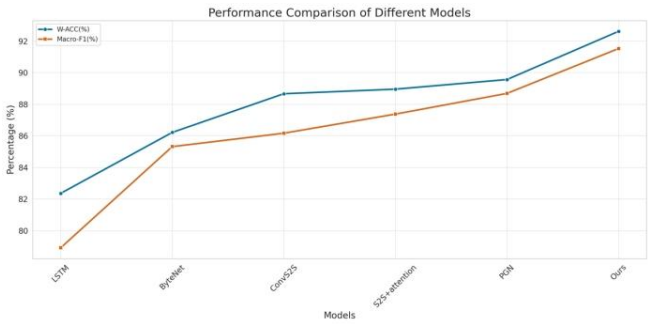

Figure 2 Evaluation results of comparison methods

B. Automatic evaluation

Table 3 demonstrates that our model outperforms comparative methods—LSTM, ByteNet, ConvS2S, S2S+attention, and PGN—across all metrics. Notably, for task-specific metrics such as CoStr, Pol, and Emp, our method registers impressive scores of 80.3%, 92.54%, and 46.4%, respectively. These represent marked improvements over the baseline LSTM by 10.62%, 8.5%, and 9.2%, substantiating the effectiveness of our model's design. Indeed, the model's capability to yield responses that are not only polite and empathetic but also strategically coherent in a counseling context is evident. The system also surpasses the PGN in these areas by margins of 2.64%, 3.4%, and 2.5%, further validating our approach in crafting a responsive legal counseling dialogue system. Additionally, the proposed method shows superior performance in PPL at 1.82 and R-LEN at 18.96, outdoing the baseline LSTM by 3.1 and 3.74 points, respectively. This suggests that our model's task specificity and contextual coherence drive it to foster a connection with the interlocutor, thereby generating contextually apt and fluid responses. Consequently, our method facilitates the production of interactions that are engaging and interactive. Furthermore, the table reveals that our model's superiority to PGN confirms the necessity of task-specific design in formulating fluent, coherent, and empathetic responses, deeply rooted in a suitable counseling strategy.

Table 3 Result of automatic evaluation for comparison methods

Model	CoStr	Pol	Emp	PPL	R-LEN
LSTM	71.25%	84.2%	38.6%	4.82	15.22
ByteNet	73.76%	85.6%	40.5%	4.51	15.89
ConvS2S	72.69%	85.1%	41.3%	3.95	16.76
S2S+attention	76.89%	87.8%	43.6%	3.23	16.93
PGN	79.23%	89.3%	45.3%	2.51	17.73
This study	81.87%	92.7%	47.8%	1.72	18.96

C. Human evaluation

Table 4 shows the human evaluation results. It can be observed that the proposed method yields better scores in terms of Con, Pol, Emp, Const, Fluen and N-Rep with a difference of 1.72, 1.17, 1.01, 1.38, 0.75, and 1.55, respectively as compared to the the baseline LSTM scores of Const: 4.36, Fluen: 4.73, and N-Rep: 4.89, which implies that contextual-coherence and fluency in our model have played a crucial role in generating consistent, fluent and non-repetitive utterances. Further, in terms of Con, Pol and Emp, our proposed method attains well scores of 4.06, 4.82, and 2.98, respectively. Consequently, it can be inferred that our proposed method is able to build a rapport with the victim, by generating engaging and interactive responses.

Table 4 presents the results of human evaluations, indicating that our proposed method outperforms the baseline LSTM model across several metrics. Specifically, the proposed method achieves superior scores in Con, Pol, Emp, Const, Fluen, and N-Rep, with respective improvements of 1.72, 1.17, 1.01, 1.38, 0.75, and 1.55. These scores reflect the pivotal role of contextual coherence and fluency in our model, which contributes to the generation of utterances that are consistent, fluid, and non-repetitive. Additionally, our method registers commendable scores of 4.06 for Con, 4.82 for Pol, and 2.98 for Emp. These findings suggest that the proposed approach effectively fosters a connection with individuals seeking counsel, as evidenced by its capacity to elicit engaging and interactive responses, thus establishing a supportive dialogue environment.

Table 4 Result of human evaluation for comparison methods

Model	Con	Pol	Emp	Const	Fluen
LSTM	2.34	3.65	1.97	2.98	3.98
ByteNet	2.45	3.57	2.11	3.17	4.05
ConvS2S	2.63	3.82	2.26	3.24	4.25
S2S+attention	3.16	3.97	2.43	3.68	4.48
PGN	3.42	4.52	2.58	4.13	4.56
This study	4.06	4.82	2.98	4.36	4.73

VI. Conclusion

This study introduces an advanced AI system for legal consultation, leveraging deep learning methodologies and a comprehensive legal knowledge base to notably improve consultation precision. The system boasts a user-friendly interface, a robust knowledge processing and inference engine, and an elaborate module for learning and optimization, all of which converge to furnish users with prompt and precise legal advice. A distinctive aspect of our system is its capacity to transition from a mere consultative role to an active participant in learning and enhancement. Utilizing cutting-edge deep learning techniques paired with the UCB optimization strategy, the system iteratively refines its performance based on user input, thus progressively elevating the quality of its services. The potential for further development is manifold, including refining the system's understanding and processing of user inquiries, enhancing the accuracy of legal knowledge application, and improving the assessment and refinement of consultation outcomes. While the current iteration focuses on legal consultation, the prospective extension of this AI application into other legal domains—such as automated legal document generation and legal management—is an exciting avenue for future exploration.

VII. References

[1] De Fuentes and R. Porcuna, "Main drivers of consultancy services: A meta-analytic approach," in J. Bus. Res., vol. 69, no. 11, pp. 4775-4780, 2016.

[2] S. Villata et al., "Thirty years of artificial intelligence and law: the third decade," in Artif. Intell. Law, vol. 30, no. 4, pp. 561-591, 2022.

[3] S. Nath et al., "New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology," in Br. J. Ophthalmol., vol. 106, no. 7, pp. 889-892, 2022.

[4] George and T. Walsh, "Artificial intelligence is breaking patent law," in Nature, vol. 605, no. 7911, pp. 616-618, 2022.

[5] M. Zhao et al., "An effective context‐focused hierarchical mechanism for task‐oriented dialogue response generation," in Comput. Intell., vol. 38, no. 5, pp. 1831-1858, 2022.

[6] M. H. P. Rizi and S. A. H. Seno, "A systematic review of technologies and solutions to improve security and privacy protection of citizens in the smart city," in Internet Things, vol. 20, p. 100584, 2022.

[7] Brożek, M. Furman, M. Jakubiec, and B. Kucharzyk, "The black box problem revisited. Real and imaginary challenges for automated legal decision making," in Artif. Intell. Law, pp. 1-14, 2023.

[8] K. M. Nor et al., "Abnormality detection and failure prediction using explainable Bayesian deep learning: Methodology and case study with industrial data," in Mathematics, vol. 10, no. 4, p. 554, 2022.

[9] D. Khurana et al., "Natural language processing: State of the art, current trends and challenges," in Multimedia Tools Appl., vol. 82, no. 3, pp. 3713-3744, 2023.

[10] Garivier et al., "KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints," in J. Mach. Learn. Res., vol. 23, no. 1, pp. 8049-8114, 2022.

[11] S. T. Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," in Array, vol. 14, p. 100157, 2022.

[12] J. Xie et al., "A transformer-based approach combining deep learning network and spatial-temporal information for raw EEG classification," in IEEE Trans. Neural Syst. Rehabil. Eng., vol. 30, pp. 2126-2136, 2022.

[13] M. R. Islam et al., "Explainable transformer-based deep learning model for the detection of malaria parasites from blood cell images," in Sensors, vol. 22, no. 12, p. 4358, 2022.

[14] Jamali et al., "A deep learning framework based on generative adversarial networks and vision transformer for complex wetland classification using limited training samples," in Int. J. Appl. Earth Obs. Geoinf., vol. 115, p. 103095, 2022.

[15] Sharma and S. Kumar, "Ontology-based semantic retrieval of documents using Word2vec model," in Data Knowl. Eng., vol. 144, p. 102110, 2023.

[16] R. Elshamy et al., "Improving the efficiency of RMSProp optimizer by utilizing Nestrove in deep learning," in Sci. Rep., vol. 13, no. 1, p. 8814, 2023.

[17] D. Irfan, T. S. Gunawan, and W. Wanayumini, "COMPARISON OF SGD, RMSProp, AND ADAM OPTIMATION IN ANIMAL CLASSIFICATION USING CNNs," in Proc. Int. Conf. Inf. Sci. Technol. Innov. (ICoSTEC), vol. 2, no. 1, pp. 45-51, Feb. 2023.

[18] K. Mishra, P. Priya, and A. Ekbal, "Help Me Heal: A Reinforced Polite and Empathetic Mental Health and Legal Counseling Dialogue System for Crime Victims," in Proc. AAAI Conf. Artif. Intell., vol. 37, no. 12, pp. 14408-14416, June 2023.

[19] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," in Decis. Anal. J., vol. 3, p. 100071, 2022.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[21] J. Gehring et al., "Convolutional sequence to sequence learning," in Proc. Int. Conf. Mach. Learn., pp. 1243-1252, PMLR, July 2017.

[22] A. Agrawal and P. Shukla, "Context Aware Automatic Subjective and Objective Question Generation using Fast Text to Text Transfer Learning," in Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 4, 2023.

[23] U. V. Ucak et al., "Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments," in Nature Commun., vol. 13, no. 1, p. 1186, 2022.

[24] W. Wang et al., "Understanding and improving sequence-to-sequence pretraining for neural machine translation," arXiv preprint arXiv:2203.08442, 2022.

[25] M. A. I. Talukder et al., "Bengali abstractive text summarization using sequence to sequence RNNs," in 2019 10th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), pp. 1-5, IEEE, July 2019.

[26] L. Sehovac and K. Grolinger, "Deep learning for load forecasting: Sequence to sequence recurrent neural networks with attention," in IEEE Access, vol. 8, pp. 36411-36426, 2020.

[27] W. Jiang et al., "Improving neural text normalization with partial parameter generator and pointer-generator network," in ICASSP 2021-2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 7583-7587, June 2021.

[28] Sage et al., "End-to-end extraction of structured information from business documents with pointer-generator networks," in Proc. Fourth Workshop Struct. Predict. NLP, pp. 43-52, Nov. 2020.