WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 1, Number 1, Year 2023
DOI: https://doi.org/10.62051/ijcsit.v1n1.18
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Exploring New Frontiers of Deep Learning in Legal Practice: A Case Study of Large Language Models

Yixu Wang[1] *, Wenpin Qian[2], Hong Zhou[3], Jianfeng Chen[4], Kai Tan[5]

[1] Computer Technology, Peking University, Beijing, China
[2] Information Science, Trine University, Phoenix AZ, USA
[3] Computer Technology, Peking University, Beijing, China
[4] Statistics, Independent Researcher, Fairfax, Va, USA
[5] Electrical & Computer Engineering, University of Washington, Seattle, WA, USA
*Corresponding Author: Yixu Wang (Email: yixu1248@gmail.com)

ABSTRACT

The LLM wave brought by ChatGPT has swept various vertical fields. Medical, finance, finance and other fields have gradually had their own exclusive large models, such as BloombergGPT, herbal medicine, Huatu, ChatMed, etc. In the legal field, we have seen LawGPT and Lawyers-LLAMA two preliminary open source models. People often think that by using domain-specific knowledge to fine-tune the model, you can get satisfactory results, but the legal field, because of its inherent requirements for accuracy, simply fine-tuning with some legal dialogue data is not enough to support the needs of real legal scenarios. Therefore, the large language model is obviously the most active field of AI at present, open source/closed source models continue to emerge, new research papers emerge in an endless stream, but as practitioners, how to truly understand the capabilities and limitations of the large language model, how to apply the large language model? The present has long been worth thinking deeply. In this paper, we propose a large language model for open source law called ChatLaw.

KEYWORDS

Large language model; ChatLaw; Legal text; Deep learning.

1. INTRODUCTION

Large Language model (LLM) refers to the deep learning model that can process large amounts of natural language data, and it has shown great potential in many fields such as natural language processing, text generation, and machine translation. In the last few years, the LLM field has experienced rapid development, with Google and OpenAI as the two leading companies in this field. Google is a significant player in the LLM space, with its BERT self-coding model and T5 codec achieving excellent performance on natural language understanding tasks. By pre-training large-scale text data, BERT model can extract word vectors and learn context information. On the basis of BERT, T5 model further integrates generative tasks into it and realizes integrated natural language processing capability. The emergence of these models has greatly promoted the development of the LLM field. In contrast, OpenAI has been using decoder only's GPT model since 2018, practicing the "violence aesthetics" - to achieve AGI with a large model path. By pre-training massive corpus data, GPT model learns the laws and patterns in natural language, and achieves excellent performance in generative tasks. OpenAI firmly believes that if the model size is large enough, a pure decoder model can achieve the goal of AGI. But because of the rapid development of the field, it is difficult to identify the

challenges that remain and the application areas that are already producing results. Expanding the amount of pre-training data has become one of the main drivers for providing LLM general capabilities. The size of the pre-training dataset quickly exceeded the number of documents that most human teams could manually quality check. Instead, most data collection processes rely on heuristic rules about data sources and filtering.

2. RELATED WORK

In the past, NLP developers relied on technology stacks such as text classification, named entity recognition, and named entity disambiguation to optimize NLP tasks. However, with the rapid development of large language models (LLMS), new technology stacks are beginning to emerge to support and accelerate the implementation and application of these large language models.

Let's follow the Langchain developers and discuss the changes taking place in the LLM and NLP technology stacks, and what these changes mean for developers.

2.1. Deep Learning and LLMs

Since the fall of 2022, a new technology stack designed to fully tap the LLM's potential has begun to emerge. In contrast to previous technology stacks, this one focuses on implementing text generation - a task that modern LLMS excel at compared to earlier machine learning models. This new technology stack consists of four main parts: data preprocessing pipeline, embeddings endpoint + vector store, LLM endpoints and LLM programming framework (LLM) programming framework).

There are several big differences between the old technology stack and the new one:

1. The new stack relies less on knowledge graphs that store structured data (such as triples) because LLMS such as ChatGPT, Claude, and Flan T-5 encode more information than earlier models such as GPT 2.

2. Newer technology stacks use off-the-shelf LLM endpoints as models, rather than custom ML pipelines (at least initially). This means that today's developers spend less time training specialized information extraction models (such as named entity recognition, relationship extraction, and sentiment analysis) and can launch solutions in less time (and cost).

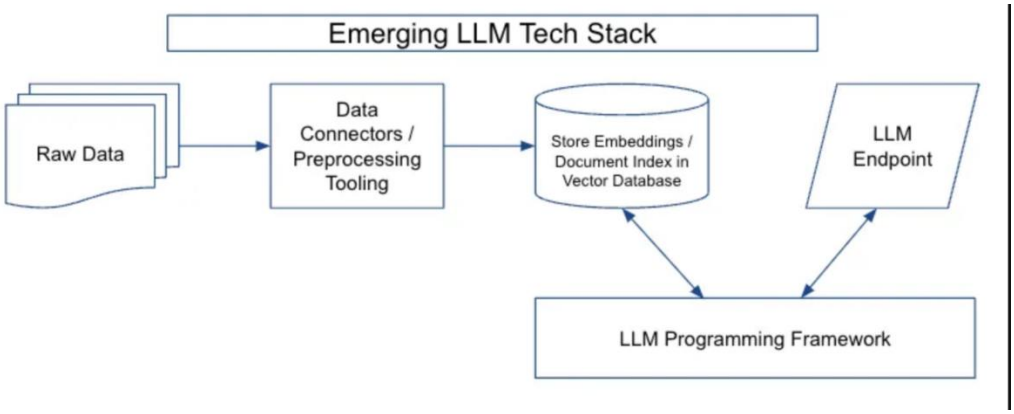

Figure 1. Emerging LLM Tech Stack

The first major part of the new stack (the data preprocessing pipeline) doesn't change much from the old stack. This step includes connectors for ingesting data (e.g. S3 bucket and CRM), data transformation layer, and downstream connectors (e.g., to vector databases); The use of embeddings endpoints and vector stores represents a significant evolution in the way data is stored and accessed. Previous embeddings were mainly used for proprietary tasks such as document clustering; The third core component of the new technology stack is the LLM terminal. This is the terminal that receives

the input data and produces the LLM output. LLM terminals are responsible for managing the model's resources, including memory and computation, and provide a scalable and fault-tolerant interface to provide LLM output to downstream applications; The last major piece of the new technology stack is the LLM programming framework. These frameworks provide a set of tools and abstractions for building applications using language models. At LangChain, this is exactly the framework we are trying to build.

2.2. ChatLaw LLM

ChatLaw=ChatLaw LLM + keyword LLM + laws LLM.

Is based on OpenLLAMA, a commercially viable model, by expanding the Chinese vocabulary and combining training data from sources such as MOSS. This enables the creation of a basic Chinese model. We then combined law-specific data to train our legal model.

Chatlaw is a legal artificial intelligence platform based on deep learning, which can help legal practitioners and ordinary users solve various legal problems and improve the efficiency of legal office and the quality of legal services. The core technology of Chatlaw is the Chinese law Grand Model, which is a natural language processing model based on the structured corpus of 100 million legal fields and professional manual annotation for knowledge injection, with powerful understanding, analysis, generation and dialogue capabilities. You can ask it for legal advice, and it looks up legal entries on its own and gives reasonable advice.

Chatlaw currently provides the following features:

(1) Structured extraction: support reading files, recordings, processing a variety of complex scenarios, one-click automatic extraction of key information, such as parties, causes, claims, factual evidence, etc. Support to summarize and sort out key facts, generate maps, charts, visual analysis.

(2) Automated writing: The writing model is fine-tuned based on the corpus of millions of legal documents, accurately summarizes user facts with one click, and automates the writing of legal documents, such as complaint, defense, judgment, etc. Supports a variety of formats and styles of output, such as Word, PDF, Markdown, etc.

(3) Intelligent question and answer: The question and answer model is trained based on a large number of legal knowledge bases and case databases, and supports accurate, professional and friendly answers to various legal questions, such as interpretation of legal provisions, case analysis, and judgment of rights and obligations. Supports multiple rounds of conversation and context understanding, providing relevant references and links.

(4) Malleable discussion: The discussion model is trained on a large number of legal topics and comments to support in-depth, interesting and insightful discussions on malleable legal topics, such as legal hot topics, industry status, career development, etc. Support multi-angle and multi-level viewpoint expression, improve users' critical thinking ability and communication ability.

Therefore, this paper proposes a method to solve the illusion problem on the experimental model published by the Peking University team, that is, strengthen the training process of the model, and add four modules in the reasoning process: "consultation", "reference", "self-suggestion" and "response". By integrating the vertical model and the knowledge base through reference modules, we inject domain-specific knowledge into the model and utilize accurate information from the knowledge base to reduce the occurrence of hallucinations.

3. CASE AND METHODOLOGY

3.1. Case

When building the dataset, we used a variety of methods to ensure the comprehensiveness and diversity of the dataset. The data set is composed as follows:

Large amounts of raw legal data: This includes the collection of legal news, social media content, and discussions in legal industry forums. These sources provide a variety of authentic legal texts that provide insight into a variety of legal topics and discussions, with examples below:

Ouestion: Written in a college student's diary full of words "must kill B, there are many ways to use, preferably poison", but no action has been taken, and one day the counselor saw the diary and reported it, is this behavior intentional homicide?

Answer: According to the information you provided, a college student wrote in his diary the words "we must kill B, there are many ways we can use, the best is poisoning", but there was no action. One day, the counselor saw A's diary and was very worried and reported the case. According to China's Criminal Law, this kind of behavior is an expression of criminal intent and innocence.

The rationale is that by incorporating data from these different sources and construction methods, our datasets cover a wide range of legal contexts, ensuring that the models developed can effectively understand and process a wide range of legal scenarios.

3.2. Key words LLM

By combining verticals specific legal knowledge base with knowledge management to create ChatLaw products, it is essential to retrieve relevant information from the knowledge base based on user queries. Therefore, it is necessary to embed the pre-trained BERT model, and then use Faiss and other methods to calculate the cosine similarity and extract the first k laws and regulations related to user queries. However, this approach tends to produce sub-optimal results when the user's problem is ambiguous. Therefore, the final goal is to extract the key information from the user query and use the vector embedding of this information to design an algorithm to improve the accuracy of matching.

Because large models have significant advantages in understanding user query mask, we fine-tune the LLM to extract keywords from user queries.

Algorithm 1 Legal retrieval based on Large Langu Model keyword extraction

Initialize the BERT model for embedding and keyword extraction model.
Initialize the legal database as $\mathcal{L}$, where $\mathbf{l}_i \in \mathcal{L}$ and i represents the i-th law. Let M be the number of laws in the legal database.
Initialize the legal scores as $\mathcal{S}$, where $s_i \in \mathcal{S}$ represents the score corresponding to the i-th law, all initialized to 0. The number of elements in $\mathcal{S}$ is also M.
Extracting keywords from user queries using a keyword extraction model, and then inputting each keyword into a BERT model to obtain a collection of $\mathcal{K}$ keyword vectors, where $\mathbf{k}_i$ represents the vector for the ith keyword, with a total of N keywords. We obtain $\mathbf{s}$ by inputting the user's question into BERT.
Initialize α for assigning weight to $\mathbf{s}$.
for i *to* N **do**
  $\mathbf{v}_i = \frac{\mathbf{k}_i}{||\mathbf{k}_i||} + \alpha \frac{\mathbf{s}}{||\mathbf{s}||}$
  for j *to* M **do**
    $s_j \leftarrow s_j + cossim(\mathbf{v}_i, \mathbf{l}_j)$
  end for
end for
return $TopK(\mathcal{S})$

After obtaining the required legal keywords through the above algorithm, the next keyword data analysis is needed. The data set in this paper is the national judicial examination questions in the past ten years, and contains the test data set of 2000 questions and their standard answers, so as to measure the model's ability to process legal multiple choice questions.

3.3. Model result

An LLM model was trained using a dataset containing case law examples from 937k countries to extract corresponding legal provisions and judicial interpretations from user queries. This legal LLM model is an important part of the ChatLaw product.

Model	Score
ChatLaw(13B)	1733.85
gpt-4	1712.03
lawyer-llama(13B)	1597.18
gpt-3.5-turbo	1573.35
OpenLLaMA(13B)	1475.55
LawGPT(7B)	1452.35

Figure 2. ELO Rankingup until June 25

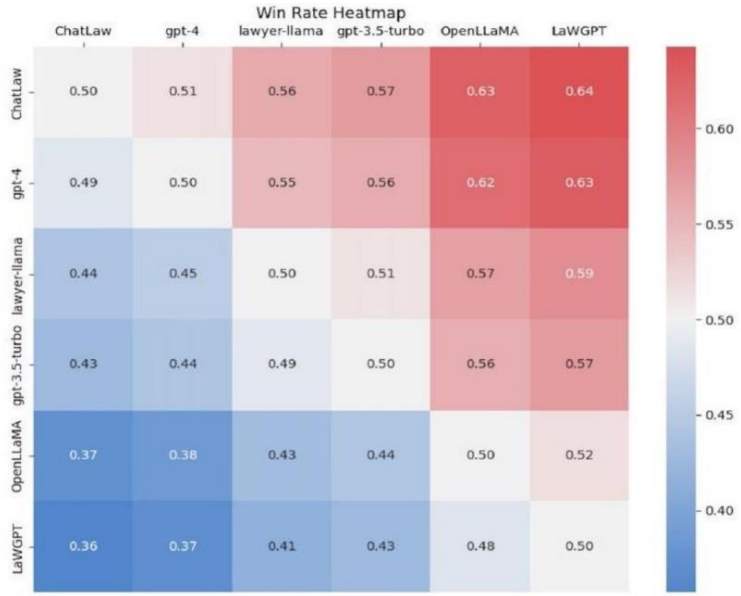

Figure 3. LLM Win Rate

Through the analysis of the above experimental results, the following conclusions can be drawn

(1) The introduction of questions and answers related to laws and regulations can improve the model's performance on multiple choice questions to a certain extent;

(2) Adding training for specific task types can significantly improve the model's performance on such tasks. For example, the ChatLaw model is superior to GPT-4 because we use a large number of multiple choice questions as training data;

(3) Multiple choice law requires complex logical reasoning, so models with more parameters usually perform better.

Based on the chatlaw algorithm developed by the team of Peking University, this paper analyzes the advantages of large language model in legal text processing in practical application. ChatLaw, a legal

large language model (LLM) developed by using knowledge in the legal field, can be seen in this paper. Based on a new method that combines LLM with vector knowledge database, this method greatly alleviates the hallucination problem common in LLM. Our stable model processing strategy addresses a wide range of legal areas. In addition, the team published a dataset of multiple choice law questions and designed an ELO model ranking mechanism.

However, limitations of this study arise from the size of the underlying model. Our performance on tasks such as logical reasoning and deduction is not ideal. In addition, after integrating a large amount of domain-specific data, further research is needed to improve the generalization of ChatLaw for general tasks. ChatLaw has potential social risks.

4. CONCLUSION

In summary, the rise of Large Language Models (LLMs) has left an indelible mark on various industries, including the legal sector. Models like ChatLaw, an open-source legal LLM, epitomize the ongoing efforts to tailor these language models for domain-specific applications. While BERT and T5 have excelled in natural language understanding, the legal field's demand for precision necessitates a more nuanced approach.

ChatLaw, situated within the evolving technology stacks discussed in related work, aligns with the industry shift toward prioritizing text generation tasks where modern LLMs excel. This model integrates legal knowledge, advanced data preprocessing, and a sophisticated programming framework, offering a comprehensive solution for structured extraction, automated writing, intelligent Q&A, and malleable discussions in legal scenarios. The case study on ChatLaw demonstrates its efficacy in processing legal multiple-choice questions and performance improvements through task-specific training. However, the study acknowledges limitations tied to model size, calling for continued research to enhance generalization. Despite these constraints, ChatLaw represents a significant stride in leveraging large language models for nuanced legal text processing, offering a promising avenue for more efficient and accurate legal services.

As the legal field grapples with the challenges and opportunities presented by AI technologies, ChatLaw stands as a notable advancement, emphasizing the imperative of understanding both the capabilities and limitations of large language models in practical applications. This progress sets the stage for further developments, fostering a symbiotic relationship between AI advancements and the complex demands of the legal profession.

ACKNOWLEDGEMENT

At the end of the article, I would like to extend my heartfelt thanks to Yu et al for their work on Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text The excellent work done in the article Processing Method. This article has provided profound insights for my research, especially in the core areas of ChatLaw, large language models, and deep learning. With the help of this paper, I can better understand and apply deep learning techniques, especially in dealing with semantic similarity matching of patent documents. The author's innovative work in the proposed integrated BERT model and novel text processing approach provides strong support for my research. In addition, I would like to thank this journal for providing a platform for the academic community to share and learn. Journals provide me with the opportunity to deeply study and discuss the application of large language model in patent document processing, and provide rich literature resources for my research. I am deeply grateful to the journal editors and all the authors involved in this work.

Article link: arXiv preprint arXiv:2401.06782.

REFERENCES

[1] Zheng He, et al. “The Importance of AI Algorithm Combined With Tunable LCST Smart Polymers in Biomedical Applications”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 92-95, https://doi.org/10.54097/d30EoLHw.

[2] S. Tianbo, H. Weijun, C. Jiangfeng, L. Weijia, Y. Quan and H. Kun, "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition," 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2023, pp. 834-837, doi: 10.1109/ICCECE58074.2023.10135464.

[3] Zhang, Quan, et al. “Deep Learning Model Aids Breast Cancer Detection”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 1, Dec. 2023, pp. 99-102, https://doi.org/10.54097/fcis.v6i1.18.

[4] Jingyu Xu, Yifeng Jiang, Bin Yuan, Shulin Li, Tianbo Song，Automated Scoring of Clinical Patient Notes using Advanced NLP and Pseudo Labeling，arXiv preprint arXiv:2401.12994, 2024

[5] Xiaonan Xu, Bin Yuan, Yongyao Mo, Tianbo Song, Shulin Li, Curriculum Recommendations Using Transformer Base Model with InfoNCE Loss And Language Switching Method, arXiv preprint arXiv:2401.09699

[6] “A Deep Learning-Based Algorithm for Crop Disease Identification Positioning Using Computer Vision”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 85-92, https://doi.org/10.62051/ijcsit.v1n1.12.

[7] “Implementation of Computer Vision Technology Based on Artificial Intelligence for Medical Image Analysis”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 69-76, https://doi.org/10.62051/ijcsit.v1n1.10.

[8] “Enhancing Computer Digital Signal Processing through the Utilization of RNN Sequence Algorithms”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 60-68, https://doi.org/10.62051/ijcsit.v1n1.09.

[9] Dong, Xinqi, et al. “The Prediction Trend of Enterprise Financial Risk Based on Machine Learning ARIMA Model”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 65-71, doi:10.53469/jtpes.2024.04(01).09.

[10] Tan, Kai, et al. “Integrating Advanced Computer Vision and AI Algorithms for Autonomous Driving Systems”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 41-48, doi:10.53469/jtpes.2024.04(01).06.

[11] “A Deep Learning-Based Algorithm for Crop Disease Identification Positioning Using Computer Vision”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 85-92, https://doi.org/10.62051/ijcsit.v1n1.12.

[12] Wang, Sihao, et al. “Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 58-64, doi:10.53469/jtpes.2024.04(01).08.

[13] Wei, Kuo, et al. “Strategic Application of AI Intelligent Algorithm in Network Threat Detection and Defense”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 49-57, doi:10.53469/jtpes.2024.04(01).07.

[14] “Based on Intelligent Advertising Recommendation and Abnormal Advertising Monitoring System in the Field of Machine Learning”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 17-23, https://doi.org/10.62051/ijcsit.v1n1.03.

[15] Yu, Liqiang, et al. “Research on Machine Learning With Algorithms and Development”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 7-14, doi:10.53469/jtpes.2023.03(12).02.

[16] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[17] Yu, L., Liu, B., Lin, Q., Zhao, X., & Che, C. (2024). Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method. arXiv preprint arXiv:2401.06782.

[18] Huang, J., Zhao, X., Che, C., Lin, Q., & Liu, B. (2024). Enhancing Essay Scoring with Adversarial Weights Perturbation and Metric-specific AttentionPooling. arXiv preprint arXiv:2401.05433.

[19] Tianbo, Song, Hu Weijun, Cai Jiangfeng, Liu Weijia, Yuan Quan, and He Kun. "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition." In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 834-837. IEEE, 2023.DOI: 10.1109/mce.2022.3206678

[20] Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science, 3(12), 36–42. https://doi.org/10.53469/jtpes.2023.03(12).06

[21] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[22] Zhang, Yufeng, et al. "Manipulator Control System Based on Machine Vision." International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019: Applications and Techniques in Cyber Intelligence 7. Springer International Publishing, 2020.

[23] Gao, Longsen, et al. "Autonomous Multi-Robot Servicing for Spacecraft Operation Extension." 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023.