
# LegalGPT: Legal Chain of Thought for the Legal Large Language Model Multi-agent Framework

Juanming Shi[1], Qinglang Guo[2], Yong Liao[3(✉)], and Shenglin Liang[2]

[1] University of Science and Technology of China, HeiFei, China
sjm2022@ustc.edu.cn
[2] China Academic of Electronics and Information Technology, BeiJing, China
gql1993@mail.ustc.edu.cn
[3] Xidian University, XiAn, China
yliao@ustc.edu.cn

**Abstract.** This paper delves into the application of Large Language Models (LLMs) in the field of legal task processing with a specific focus on the emerging capabilities these models have demonstrated in complex reasoning and zero-shot learning (ZSL). By introducing a multi-intelligence framework based on Large Scale Legal Language Modeling, the research aims to improve the efficiency and performance of the models across a wide range of functions including legal review, consultation, and judicial decision-making. The framework utilizes function-specific legal chains of thoughts and specialized agents to guide the LLMs to handle legal tasks, optimize response behaviors, and enhance reasoning capabilities. The study also incorporates an information retrieval module to address the common hallucination problem of LLMs, thereby improving response reliability and enhancing the model's ability to tackle complex legal issues. The evaluation of the LegalGPT model shows that it indeed outperforms the existing legal LLMs in terms of accuracy, completeness, and linguistic quality in several Chinese legal domains.

**Keywords:** LegalGPT · Large Language Model · chain of thought · Agent

## 1 Introduction

Large Language Models [1–7] have achieved significant advances across a wide range of language comprehension tasks, advancing the state-of-the-art in language comprehension capabilities. The successful application of these models have led to their widespread deployment and adoption across a variety of programs [8–12]. In particular, LLMs demonstrate emerging capabilities, such as complex reasoning [13–15], which have rendered them a prominent research topic in recent years. This capability is readily explicable because LLMs can learn through intermediate reasoning steps generated from a minimal number of samples. Furthermore, research has unveiled the ability of LLMs to perform zero-shot learning (ZSL) without any task-specific examples [16, 17]. The zero-shot learning capability of LLMs substantially reduces the cost of data annotation when compared with supervised learning or fine-tuning. This inspires us to explore the implementation of LLMs and thought chains in the field of AI and law.


D.-S. Huang et al. (Eds.): ICIC 2024, LNAI 14880, pp. 25–37, 2024.
https://doi.org/10.1007/978-981-97-5678-0_3

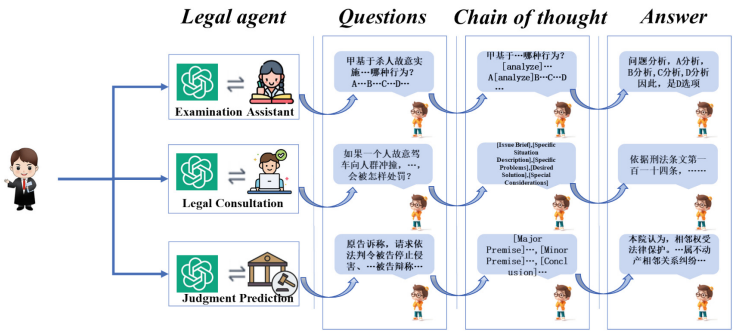


**Fig. 1.** Three types legal tasks from top to bottom: legal examinations, legal consultation, and judicial decisions. When a user poses a question, the model employs the appropriate one of these three legal chain of thought functions to analyze the question and ultimately generate an answer.

In this study, we categorize tasks within the legal domain into three primary types: legal examinations, legal consultation, and judicial decisions, as depicted in Fig. 1. For each task type, we developed distinct specialized agents tasked with guiding the reasoning process of LLMs in addressing legal tasks [18, 19]. Simultaneously, to align the reasoning process of LLMs more precisely with the requirements of each specific task, we formulated the distinct legal thought chains for each type of legal task. Additionally, we engineered a further agent specifically devised to guide LLMs in generating instructions pertaining to legal thought chains [19–22]. The design philosophy of this approach is underpinned by a combination of two lines of work. Initially, linguistic agent technologies were developed to automate specific tasks. In our study, LLMs in the legal domain are utilized to generate instructions that satisfy the user's needs, with legal agents not only completing the task but also providing specific instructions for its completion. This is achieved by augmenting and replicating manual work through autonomous agents for LLMs: initially, we collect the relevant data needed for a specific legal function, followed by the supervised fine-tuning of the large model. The agent identifies the desired legal function based on user input and routes the input to a specific model for further inference. Thus, our approach adheres to the latest concepts of agent design, by integrating agents with LLMs, utilized to generate plans for processing flows. Secondly, chain-of-thought (COT) reasoning has produced significant results in tasks such as arithmetic reasoning [16, 23]. In the field of Natural Language Processing (NLP), prompting strategies are frequently utilized in conjunction with small-sample or zero-shot learning [13, 24], aimed at optimizing and tuning the response behavior of LLMs to achieve higher accuracy across a variety of complex NLP scenarios. Under these specific learning frameworks, LLMs are trained with a series of carefully crafted example cues that direct the model toward deeper conceptual generalizations, enabling efficient performance across a range of NLP tasks. Cue learning equips LLMs with precise and informative instructions, often manifested as targeted questions or statements, designed to steer the model toward producing outputs consistent with legal knowledge. The technique of "thought chain cueing" is applicable when addressing complex tasks such as legal quizzing, judgment processing, and case retrieval. This technique encourages the

model to undertake more complex legal reasoning[25] or multi-stage decision-making processes through a series of strategically designed and interconnected prompts [21].

In summary, this study did some investigations in the following aspects: (1) A multi-intelligence framework based on LLMs has been developed. This framework is a high effective platform to integrate multi-intelligence components or various functions. (2) Three chains of function-specific legal thought have been formulated. These COT improves the collaborations among LLMs-based agents, their utilization leads to the higher robustness and efficiency of the whole system. (3) An information retrieval module(e.g., for the extraction of legal terms) has been implemented to mitigate the hallucination issue [26] and produce more reliable responses. This application of this module could effectively eliminate irrelevant information from the external knowledge base and increases the reasoning capabilities of the legal LLMs.

## 2 Relate Works

In this section, a comprehensive review has been implemented for the legally relevant chains of thought and the content of the Large Language Model-based Agent.

### 2.1 Language Agent

Recent advancements have seen autonomous agents driven by Law Large Models (LLMs) attract significant interest across industry and academia [21]. These studies have improved LLM problem-solving capabilities by facilitating discussions among multiple agents, stabilizing alignment values, creating instructional datasets, and analyzing social dynamics and collective memory in agent societies [23, 27–30]. Furthermore, "brainstorming" sessions among functionally diverse agents have been used to tackle complex tasks, alongside cost-reduction strategies employing varying model sizes [31].

Our team has developed a multi-intelligence LLM framework tailored specifically for the legal domain. This framework assists with legal exams, advice, and case adjudication, diverging from conventional studies on agent cooperation or competition by focusing on delivering precise legal services. It addresses common challenges in legal scenarios, such as "auxiliary repetitive instructions" and "infinite message loops", by incorporating specific legal thought chains and customizing the framework for legal prompts, thus enhancing collaboration in legal practice, research, and education. This approach not only seeks to enhance legal service efficiency and accuracy but also aims to advance legal technology and provide valuable insights for other complex and specialized fields.

### 2.2 Large Language Models and Chain of Thought

LLMs have become instrumental in predicting word distribution within contexts by training on vast textual data, demonstrating high efficiency in various NLP tasks through simple cueing techniques. Building on this, Wei et al. introduced the Chain of Thought (COT) methodology, which guides the model to answer mathematical questions step-by-step using brief hints [13]. Additionally, Kojima et al. developed Zero-Shot COT, employing the hint "Let's think step by step" to unlock advanced reasoning capabilities

in models [16]. This methodology was extended by Wei et al., who explored multi-step reasoning in larger models, a feature absent in smaller counterparts. Yu et al. applied Zero-Shot COT to the COLIEE legal entailment task, achieving optimal results and showcasing the potential of LLMs in legal reasoning [32].

Leveraging these insights, we've created three specialized thought chains for our Legal Large Language Model tailored to legal exams, consultation, and case adjudication. These enhancements significantly boost the model's legal reasoning abilities, making it more proficient in complex legal scenarios and marking advancements in technology application within the legal sector, as well as in advanced reasoning and decision-making strategies.

## 3 Approach

As illustrated in Fig. 2, this section denotes the processes in the COT of designing the three functionalities of LegalGPT including the Legal Examination Assistant COT(Sect. 3.1), the Legal Consultation COT (Sect. 3.2), and the Judgment Prediction COT(Sect. 3.3), the process for fine-tuning LegalGPT has been presented in Sect. 3.4.

### 3.1 Legal Examination Assistant COT

The Legal Examination Assistant COT enhances the reasoning accuracy of the Large Language Model of Law in addressing law exam questions through a series of innovative approaches. Traditional processing methods typically yield answers directly without including the question parsing and in-depth reasoning process, it may lead to inaccurate results. In contrast, as shown in Fig. 2(a), the Legal Examination Assistant COT adopts a more granular approach of processing to improve the accuracy of reasoning and answering.

First, the system meticulously analyzes each option, covering aspects such as the specific issue addressed by the option and the legal direction in which the option is designed etc.. For example, it identifies the legal direction (e.g., criminal law) primarily involved in a question and further analyzes whether there is a direct correlation leading to death. Through this approach, the system gains an in-depth understanding of the problem's nature and the options' relevance. Next, the system evaluates the legal direction of each option and its alignment with the question to ascertain which option most accurately captures the question's intent. This process incorporates not only the legal logic of the question but also a comprehensive understanding and application of legal knowledge. Additionally, the Legal Examination Assistant COT incorporates an external knowledge base and rethink mechanism for questions demanding memorized legal knowledge, such as law memorization and case retrieval. The system utilizes the external knowledge base to re-reason based on previous results and analysis, thereby correcting and optimizing answers through the rethink mechanism. This step has been demonstrated in practice to significantly improve answer accuracy. This approach not only serves as a powerful tool for legal examinations but also shows the potential of large language models for intricate reasoning and knowledge application.


**Fig. 2.** Detailed explanations of the thought chain processes underpinning the three functionalities of LegalGPT are as follows: (a) the Legal Examination Assistant COT, (b) the Legal Consultation COT, and (c) the Judgment Prediction COT.

## 3.2 Legal Consultation COT

The Legal Consultation COT employs a structured approach to legal advice, which includes the use of an external knowledge base to enhance response professionalism and aims to improve the linguistic quality and accuracy of the responses. Inspired by prior research and professional guidance, this approach largely increases the capability to understanding the questions clearly and comprehensively through a five-part method for organizing and processing legal inquiries.

The five parts are as follows:

**Issue Brief:** Outlines the primary area of advice sought by the user (e.g., labor law, contract law), and the central problem or objective the user aims to resolve.
**Specific Situation Description:** Provides details on the involved parties, the timing and location of the incident, and any actions taken, ensuring the Big Model of Law has a clear understanding of the entire situation.
**Specific Problems:** Lists specific, well-defined issues, ideally in the form of closed-ended questions that directly relate to legal texts, case studies, or professional advice.
**Desired Solution:** Describes the specific outcomes the user seeks through counseling, such as compliance, risk mitigation, obtaining compensation, and the type of advice desired, including legal counsel, operational steps, and preventive actions.
**Special Considerations:** Mentions specific time and budget constraints, or other factors that may influence the implementation of the advice.

The Legal Consultation COT model utilizes the five-part structure of the consultation questions posed by the user, combined with an external knowledge base, to generate

reasoning results and answers for the user's consultation. This structured questioning and processing approach not only guarantees clear and comprehensive understanding by the legal expert but also significantly boosts the likelihood of obtaining useful and accurate legal advice. Through this method, Legal Consultation COT effectively handles complex legal inquiries, providing users with professional, accurate, and high-quality legal advice.

### 3.3 Judgment Prediction COT

The Judgment Prediction COT model enhances the accuracy and reliability of judgment predictions through a trinomial approach to determining judgment outcomes [44]. Inspired by traditional logical reasoning structures and the capabilities of large-scale language models to process legal texts, this approach:

**Major Premise:** Following the user's input of the case, relevant legal texts are retrieved. These texts outline the legal framework and serve as the basis for judgment, ensuring compliance with existing legal provisions.
**Minor Premise:** The user's case facts are simplified to outline specific details. This transformation reduces the case's complexity to actionable facts, establishing the groundwork for subsequent logical reasoning.
**Conclusion:** Utilizing the major and minor premises, logical reasoning determines a verdict. This involves applying legal principles to the case facts to foresee the most probable outcome.

With this methodology, the Judgment Prediction COT not only conducts in-depth analysis in alignment with the legal text and case facts but also logically and systematically reasons towards a verdict. This structured reasoning enhances prediction accuracy and interpretability, enabling both legal professionals and general users to more comprehensively understand the basis of the predictions. By precisely integrating legal texts with case facts within a classical logical reasoning framework, the model serves as a valuable resource for legal professionals, while also pioneering new avenues for research and application in legal technology.

### 3.4 Legal Large Language Model Fine Tuning

In order to address real-world problems, this study emphasizes the importance of the model's ability to reasoning in the legal domain. With this in mind, we selected supervised data from downstream tasks and conducted instructive fine-tuning in the model. Given that the model is primarily designed to deal with the real-world legal counselling problems, our fine-tuning process includes three main aspects: first, utilizing public NLP legal task datasets such as JEC-QA [34], CJRC [35]; second, utilizing open-source instructional datasets including Lawyer-Llama [36] and LawGPT-zh [37]; and finally, relying on a large amount of adjudication data and integrating professional classification processing of the corpus. In this study, the Baichuan-13B-chat model served as the foundation for the specialized fine-tuning tailored to each of the three problem domains: Examination Assistant, Legal Consultation, and Judgment Prediction. In addition, given the stringent language quality standards in Legal Consultation, and to prevent overfitting

to a specific legal corpus, we further integrated generic datasets, such as gpt4_data_zh [5] and Firefly [38], into the fine-tuning process to ensure linguistic accuracy and fluency. The detailed fine tuning process has been presented in the other paper [Legal-LM: Knowledge Graph enhanced Large Language Models for Law Consulting]. Here we will skip this part. The final optimized parameters and settings are as following: PEFT technique has been used, the study rate is 5E-5 with two training cycles, the maximum length has 2048 cards, the maximum target length is 1024 cards. The setting of LoRA is $8^{th}$, Alpha is 6, the dropout of LoRA is 0.05. The data type is float16 for Torch fp16.

## 4 Experiments

In this section, to evaluate the capabilities of Legal-LM, we conducted baseline experiments that focused on assessing both the overall performance of the model and its efficacy in responding to legal advice queries and judgment prediction inquiries, followed by an in-depth analysis of the experimental outcomes.

### 4.1 Experiments Setting

**Datasets:** Employing both objectively judged datasets and subjective datasets to ensure diversity and representativeness in the assessment constitutes a thoughtful and effective strategy. This approach facilitates a comprehensive assessment of the Legal Large Language Model's performance from the different perspectives, incorporating both objective evaluations from standardized test questions and subjective evaluations from actual legal scenarios.

Objective Judging Dataset: this dataset encompasses a range of legal questions with the different level of difficulties, including: **LBK(Legal Basics Question Bank)**: [39] Of simple difficulty, focusing on basic legal knowledge, it assesses the model's grasp of fundamental legal concepts and principles. **UNGEE (National Unified Examination for the Master of Laws):** [39] Of medium difficulty, this dataset aligns closely with the specialized knowledge of legal professionals, testing the model's advanced law knowledge and application. **NJE (China's Unified Qualification Exam for Legal Professionals):** [39]With higher difficulty, it covers extensive legal knowledge and practical skills, serving as an advanced test of the model's comprehensive legal analysis and application. Evaluating these three datasets of varying difficulties provides a comprehensive assessment of the legal grand model's problem-solving abilities at different levels and its adeptness at handling both single-choice and multiple-choice questions.

The subjective dataset, manually edited by lawyers, features 1,000 examples drawn from legal advices, online forums, justice-related publications, and legal documents. It encompasses a broad spectrum of scenarios, including legal quizzes, legal consultation, and verdict prediction. This dataset construction approach aligns the assessment closely with real-world legal practices, testing the model's effectiveness in addressing actual legal issues. Such an assessment system enables a thorough evaluation of the legal grand model's capabilities in theoretical knowledge, legal reasoning, case analysis, and practical application. Combining objective and subjective datasets enriches the assessment dimensions and ensures diversity and representativeness in the results, more accurately rethinking the model's performance and application potential.

**Evaluation Protocols:** Adopting this comprehensive assessment methodology is crucial to ensuring that the evaluation of the legal grand model is both thorough and fair. By integrating the accuracy metrics from the objective dataset with the detailed scoring of the subjective component, the legal grand model's performance across various dimensions can be thoroughly evaluated to provide more accurate and representative assessment outcomes.

Objective assessment: using the accuracy rate as a metric in the objective assessment section offers a straightforward and effective method to gauge the Legal Large Language Model's proficiency in addressing standardized test questions. Accuracy effectively demonstrates the model's competency in identifying correct legal concepts and applying legal knowledge.

Subjective assessment: this utilizes a sophisticated, in-depth evaluation methodology to simulate a real law exam's Q&A process and incorporate reviews by legal professionals, to ensure professionalism and accuracy. The four qualities are:

**Accuracy:** indicates whether the answer correctly addresses the posed question, serving as the fundamental and most critical criterion for evaluating legal responses.
**Completeness:** measures the extent to which the answer encompasses all key points of the question, rethinking the respondent's comprehensive understanding and mastery of the legal issue.
**Clarity:** Evaluates the clearness of the answer's logical structure and language, crucial for showcasing the respondent's logical thinking and presentation skills.
**Language Quality:** assesses the answer's language expression, including grammar, spelling, punctuation, and the fluency and naturalness of language, as a key indicator of expressive ability.

This assessment method prioritizes the correctness of the answer, its expression and logical structure, it benefits the Big Model of Law for its capacity to tackle the complex legal issues. Combining these scoring criteria yields an evaluation of the model's comprehensive abilities in legal specialties, this evaluation mechanism guides further optimization and enhancement of the Law Grand Model to make its better adaptation to the unique demands of the legal field.

**Baseline:** To validate the effectiveness of our proposed LegalGPT design, we conducted a benchmark evaluation of legal large language models and various baseline models across different research areas. For the benchmark evaluation, we selected two types of models: The Chinese Large Language model and the Chinese law large language model. The Chinese base large language model lineup includes Ziya-LlaMA-13B [40], ChatGLM-6B [41], and Baichuan-13B-Chat [42]; the Chinese legal large language model comprises LexiLaw [43], LawGPT-7B [37], Lawyer LlaMA-13B [36], ChatLaw-13B [44], and DISC-LawLLM-13B [39].

### 4.2 Results

In order to obtain the fair estimation for the objective dataset, we used the accuracy as the key deciding index for our LLMs model's overall performance. The computation

of accuracy could be formulated as: accuracy $= \frac{d}{N}$, where N is the total number of the objective dataset, d is the number of the correct answers to the large law language model. For the subjective dataset, we created a testing dataset based on the published papers/books and other online documents related to legal consultation, legal forums and law. The questions in this testing dataset have been answered by anonymous different LLMs, then three legal personals will give a score for the four aspects of accuracy, completeness, clarity and language quality, the average value for each aspect will be used as the final criteria to decide which one is relative better (higher score means better performance).

Table 1 shows the accuracy performance of all methods across the three objective datasets. For the LBK dataset, LegalGPT's accuracy surpasses these of other models by around 15 percentage points. For the UNGEE dataset, LegalGPT's accuracy exceeds the top-performing baseline model by 16 percentage points. For the NJE dataset, LegalGPT's accuracy is at least 20 percentage points higher. The data in Table 1 clearly demonstrates that LegalGPT outperforms all baseline models across all three challenging legal problem categories. Additionally, comparing the full model with and without the rethink mechanism, the complete LegalGPT model consistently surpasses its counterpart lacking this mechanism by 5 percentage points in each dataset category.

**Table 1.** Accuracy evaluation of LegalGPT and other comparative baseline models on three experimental datasets

| Model | LBK | UNGEE | NJE |
|---|---|---|---|
| Ziya-LlaMA-13B | 43.27 | 40.94 | 25.70 |
| ChatGLM-6B | 42.91 | 39.69 | 31.66 |
| Baichuan-13B-Chat | 53.45 | 50.00 | 31.47 |
| LexiLaw | 40.36 | 31.56 | 20.11 |
| LawGPT-7B | 29.09 | 30.31 | 22.91 |
| Lawyer LlaMA-13B | 39.64 | 32.50 | 35.75 |
| ChatLaw-13B | 41.09 | 35.62 | 27.56 |
| LegalGPT(w/o rethink) | 62.12 | 59.11 | 55.09 |
| LegalGPT | 68.17 | 64.19 | 60.01 |

Figure 3 shows the comparisons of the four aspects of accuracy, completeness, clarity and language quality for the subjective datasets. We assessed the performance of ChatGLM, Baichuan, GPT-3.5-Turbo, Chatlaw, and Disclaw large language models using a subjective review set of 1,000 questions, primarily focused on legal advice and case verdict inquiries. In this comparison, LegalGPT showcased superior answering capabilities, particularly with subjective questions, significantly surpassing other baseline and comparable models in the legal domain. Specifically, LegalGPT's accuracy score surpassed the baseline performance by 10 points, marking a significant lead over competing models. For completeness, LegalGPT outperformed the other baseline models by a minimum of

10 points. Additionally, LegalGPT exceeded other models' clarity assessments by more than 10 percentage points. Lastly, regarding language quality, LegalGPT marginally surpassed the reference models. These findings underscore LegalGPT's unique advantages for the legal profession, particularly in understanding and responding to complex legal questions. Collectively, its high marks in accuracy, completeness, clarity, and language quality show LegalGPT's leadership in addressing legal advice and case decisions. These strengths equip legal professionals with an effective tool for legal consultation and adjudication, enhancing productivity and decision-making quality.

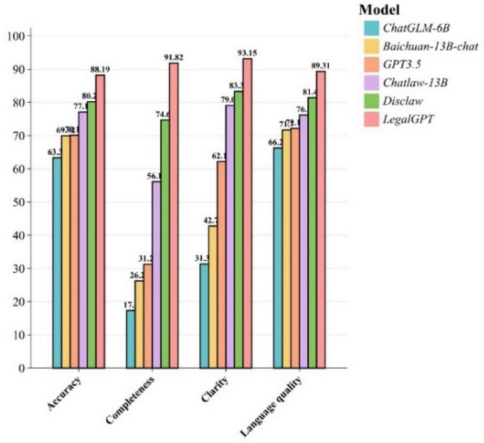


**Fig. 3.** This figure presents comparative results of Legal-LM against ChatGLM, Baichuan, Chatlaw, Disclaw, and GPT-3.5-Turbo in terms of providing subjective legal advice and making case decisions.

### 4.3 Analysis

Three main factors drive LegalGPT's superior performance. First, LegalGPT benefits from the use of extensive datasets, including law exam data like LBK, UNGEE, and NJE, which are fine-tuned to enhance performance. Second, a rethink mechanism and the integration of an external knowledge base enable the model to accurately retrieve and answer similar questions, thus boosting its capabilities. Third, the Legal Examination Assistant COT's precise analysis of questions and options significantly improves answer accuracy.

Other models often fail to surpass baseline models in the Legal Exam dataset due to baseline models' use of extensive law-related data in pre-training. In contrast, some large legal models add extra legal data and exam questions during pre-training and fine-tuning, which can lead to data illusion and contamination, reducing answer accuracy.

LegalGPT excels in handling subjective legal consultation questions due to strategic data integration during training, minimized data contamination through professional evaluation, and enhanced model performance through specialized COT approaches, improving inference accuracy, clarity, and language quality.

In subjective legal advice scenarios, the legal grand model outperforms baseline models. This success is attributed to its richer legal data pre-training, contrasting with baseline models that lack specific legal consultation scenario training. This highlights the importance of specialized data training in boosting a model's effectiveness within the legal domain.

## 5 Conclusion

This paper thoroughly explores the progress made in the application of LLMs in the legal domain, highlighting their potential for performing complex reasoning tasks and zero-shot learning. The research centers on proposing an innovative approach: the construction of a framework that integrates multiple intelligences around a core Legal Large Language Model. The framework seeks to enhance the efficiency and effectiveness of the model in handling law-related tasks, including review, consultation, and adjudication. By integrating specially-designed legal thought chains and agents, this framework substantially improves the model's legal reasoning and decision-making capabilities. The integration of the information retrieval module effectively mitigates the LLM's difficulties with hallucination problems, it results in improving the model's response accuracy and capacity to handle complex legal topics. Evaluation results for the LegalGPT model demonstrate that it markedly outperforms the existing legal language models in terms of accuracy, completeness, and linguistic quality, it proves the potential of LegalGPT in the field of legal AI.

**Acknowledgment.** This work is supported by the National Key Research and Development Program of China (2021YFC3300500); This work is also supported in part by the Natural Science Foundation of China under Grant U20B2047, U20B2053, U19B2044.

## References

1. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
2. Black, S., et al.: GPT-NeoX-20B: an open-source autoregressive language model. arXiv preprint arXiv:2204.06745 (2022)
3. Zhang, S., et al.: OPT: open pre-trained transformer Language models. arXiv preprint arXiv:2205.01068 (2022)
4. Smith, S., et al.: Using deepspeed and megatron to train megatron-turing NLG 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990 (2022)
5. OpenAI: GPT-4 Technical report. arXiv abs/2303.08774 (2023)
6. Penedo, G., et al.: The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116 (2023)
7. Anil, R., et al.: PaLM 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
8. Araci, D.: FinBERT: financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063 (2019)
9. Huang, K., Altosaar, J., Ranganath, R.: ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)

10. Wu, S., et al.: BloombergGPT: a large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
11. Driess, D., et al.: PaLM-E: an embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
12. Huang, S., et al.: Instruct2ACT: mapping multi-modality instructions to robotic actions with large language model. arXiv preprint arXiv:2305.11176 (2023)
13. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
14. Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
15. Kıcıman, E., et al.: Causal reasoning and large language models: opening a new frontier for causality. arXiv preprint arXiv:2305.00050 (2023)
16. Kojima, T., et al.: Large language models are zero-shot reasoners. Adv. Neural. Inf. Process. Syst. **35**, 22199–22213 (2022)
17. Wan, X., et al.: Better zero-shot reasoning with self-adaptive prompting. arXiv preprint arXiv:2305.14106 (2023)
18. Yao, S., et al.: ReAct: synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
19. Shinn, N., et al.: Reflexion: language agents with verbal reinforcement learning. Adv. Neural Inf. Process. Syst. **36** (2024)
20. Park, J.S., et al.: Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (2023)
21. Wang, L., et al.: A survey on large language model based autonomous agents. arXiv preprint arXiv:2308.11432 (2023)
22. Xi, Z., et al.: The rise and potential of large language model based agents: a survey. arXiv preprint arXiv:2309.07864 (2023)
23. Wang, L., et al.: Plan-and-solve prompting: improving zero-shot chain-of-thought reasoning by large language models. arXiv preprint arXiv:2305.04091 (2023)
24. Liu, P., et al.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
25. Sartor, G.: Legal reasoning. Treatise Legal Philos. Gen. Jurisprud. **5** (2005)
26. Jansen, B.J., et al.: The illusion of data validity: why numbers about people are likely wrong. Data Inf. Manage. **6**(4), 100020 (2022)
27. Chen, J., et al.: S-Agents: self-organizing agents in open-ended environment. arXiv preprint arXiv:2402.04578 (2024)
28. Zhuge, M., et al.: Mindstorms in natural language-based societies of mind. arXiv preprint arXiv:2305.17066 (2023)
29. Hao, R., et al.: ChatLLM network: More brains, more intelligence. arXiv preprint arXiv:2304.12998 (2023)
30. Liu, R., et al.: Training socially aligned language models in simulated human society. arXiv preprint arXiv:2305.16960 (2023)
31. Cai, T., et al.: Large language models as tool makers. arXiv preprint arXiv:2305.17126 (2023)
32. Yu, F., Quartey, L., Schilder, F.: Legal prompting: teaching a language model to think like a lawyer. arXiv preprint arXiv:2212.01326 (2022)
33. Jiang, C., Yang, X.: Legal syllogism prompting: teaching large language models for legal judgment prediction. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (2023)
34. Zhong, H., et al.: JEC-QA: a legal-domain question answering dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)

35. Duan, X., et al.: CJRC: a reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In: Sun, M., HUANG, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 111112019. LNCS (LNAI), vol. 11856, pp. 439–451. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_36
36. Huang, Q., et al.: Lawyer llama technical report. arXiv preprint arXiv:2305.15062 (2023)
37. Nguyen, H.-T.: A brief report on lawGPT 1.0: a virtual legal assistant based on GPT-3. arXiv preprint arXiv:2302.05729 (2023)
38. Hassanzadeh, T., Meybodi, M.R.: A new hybrid approach for data clustering using firefly algorithm and K-means. In: Proceedings of the 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012). IEEE (2012)
39. Yue, S., et al.: DISC-LawLLM: fine-tuning large language models for intelligent legal services. arXiv preprint arXiv:2309.11325 (2023)
40. Lu, J., et al.: Ziya-VL: bilingual large vision-language model via multi-task instruction tuning. arXiv preprint arXiv:2310.08166 (2023)
41. Du, Z., et al.: GLM: general language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360 (2021)
42. Yang, A., et al.: Baichuan 2: open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)
43. Dai, Y., et al.: LAiW: a Chinese legal large language models benchmark (a technical report). arXiv preprint arXiv:2310.05620 (2023)
44. Cui, J., et al.: ChatLaw: open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092 (2023)