Prolog-Driven Rule-Based Diagnostics with Large Language Models for Precise Clinical Decision Support

Xiaoyu Tan[1,2,*], Bin Li[1,*], Weidi Xu[2], Chao Qu[2], Wei Chu[2], Yinghui Xu[3], Yuan Qi[3], and Xihe Qiu[1,✉]

[1] School of Electrical and Electronic Engineering, Shanghai University of Engineering Science, Shanghai 201600, China
qiuxihe1993@gmail.com
[2] INFLY TECH(Shanghai)Co.,Ltd, Shanghai 200030, China
[3] AI^3 Institute, Fudan University, Shanghai 200082, China

Abstract. Recently, large language models (LLMs) have been increasingly utilized for decision support across various domains. However, due to their probabilistic nature and diverse learning influences, LLMs can sometimes generate inaccurate or fabricated information, a phenomenon known as “hallucination”. This issue is particularly problematic in fields like medical diagnosis, where accuracy is crucial and the margin for error is minimal. The risk of hallucination is exacerbated when patient data are incomplete or vary across different clinical departments. Consequently, using LLMs directly for clinical decision support presents significant challenges. In this paper, we introduce ProCDS, a system that integrates Prolog-based rule diagnostics with LLMs to enhance the precision of clinical decision support. ProCDS begins by converting medical protocols into a set of rules and patient information into facts. Then, we design an update cycle to extract and update related facts and rules due to possible discrepancies and missing patient information. After that, we perform a logical inference using the Prolog engine and acquire the response. If the Prolog engine cannot produce certain results, ProCDS would perform another iteration of facts and rules update to fix the potential mismatch and perform logical inference again. Through this iterative neuro-symbolic integrated process, ProCDS can perform transparent and accurate clinical decision support. We evaluated ProCDS in Obstructive Sleep Apnea Hypopnea Syndrome (OSAHS) real-world clinical scenarios and other logical reasoning benchmarks, achieving high accuracy and reliability in our results. Our project page is available at: https://github.com/testlbin/procds.

Keywords: Large Language Models · Obstructive Sleep Apnea Hypopnea Syndrome · Clinical Decision Support.

[*] These authors contributed equally to this work.
[✉] Corresponding author.

1 Introduction

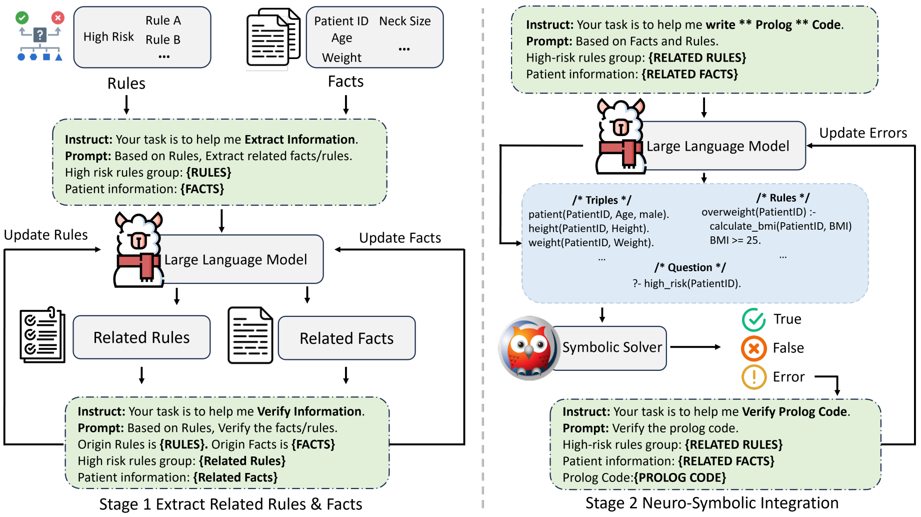
High Risk
Rule A
Rule B
...
Rules
Patient ID
Neck Size
Age
Weight
...
Facts
Instruct: Your task is to help me Extract Information.
Prompt: Based on Rules, Extract related facts/rules.
High risk rules group: {RULES}
Patient information: {FACTS}
Update Rules
Update Facts
Large Language Model
Related Rules
Related Facts
Instruct: Your task is to help me Verify Information.
Prompt: Based on Rules, Verify the facts/rules.
Origin Rules is {RULES}. Origin Facts is {FACTS}
High risk rules group: {Related Rules}
Patient information: {Related Facts}
Stage 1 Extract Related Rules & Facts
Instruct: Your task is to help me write ** Prolog ** Code.
Prompt: Based on Facts and Rules.
High-risk rules group: {RELATED RULES}
Patient information: {RELATED FACTS}
Large Language Model
Update Errors
/* Triples */
patient(PatientID, Age, male).
height(PatientID, Height).
weight(PatientID, Weight).
...
/* Rules */
overweight(PatientID) :-
calculate_bmi(PatientID, BMI)
BMI >= 25.
...
/* Question */
?- high_risk(PatientID).
Symbolic Solver
True
False
Error
Instruct: Your task is to help me Verify Prolog Code.
Prompt: Verify the prolog code.
High-risk rules group: {RELATED RULES}
Patient information: {RELATED FACTS}
Prolog Code:{PROLOG CODE}
Stage 2 Neuro-Symbolic Integration

Fig. 1. The workflow overview of our proposed ProCDS framework.

Recent advancements in large language models (LLMs) have facilitated their widespread adoption as general assistants for various daily tasks [29]. Due to their robust reasoning capabilities and fluent interaction with natural language, LLMs are increasingly deployed in specialized professional fields, including finance and healthcare [13,8]. In these domains, they offer targeted knowledge services and support for professional decision-making. However, the application of LLMs in such contexts, which have little tolerance for inaccuracies, is challenged by the propensity of these models to generate false information—a phenomenon commonly referred to as "hallucination" [11]. This issue underscores the need for enhanced LLM reliability in high-stakes environments.

Hallucination is a well-recognized issue in utilizing LLMs, stemming primarily from their probabilistic model's nature. These models are inherently prone to sampling incorrect tokens within different sampling algorithms [30], which can lead to compounded errors in their logical reasoning processes. Factors such as noise in the pre-training corpus, the usage of datasets with potentially incorrect instructions for fine-tuning, and imbalanced data distribution can further exacerbate hallucinations. Additionally, the input context can influence model generations, increasing the likelihood of hallucination due to either insufficient information or the inclusion of irrelevant data [10]. **This issue is particularly prevalent in clinical domains, where patient records often vary significantly across different departments [11]. Consequently, employing**

LLMs for reliable and trustworthy clinical decision support presents substantial challenges.

Recently, several studies have explored the use of neural-symbolic frameworks to address the problem of hallucinations in LLMs. These approaches involve converting natural language logical reasoning tasks into executable code, which can be processed by logical engines or code interpreters [27,26]. This transformation allows LLMs to generate accurate results based on logical engine-generated specific outputs, significantly reducing instances of hallucination in reasoning processes. However, this process heavily relies on prompt engineering and domain-specific knowledge to consistently translate natural language inputs into executable code. It often necessitates the explicit coding of key information to enable efficient inference [25]. **Consequently, these frameworks face substantial challenges in real-world clinical decision support scenarios, where patient information varies and may require diverse logical inference pathways for each individual.**

Hence, is it feasible to develop a clinical decision support framework that incorporates adaptive logical reasoning based on diverse patient data? In this paper, we propose a novel framework, ProCDS[4], which leverages **Pro**log-driven, rule-based diagnostics using LLMs for robust and interpretable **C**linical **D**ecision **S**upport. The ProCDS framework begins by transforming medical protocols into a structured set of rules and converting patient data into factual inputs. It includes a dynamic update cycle to refine and expand these rules and facts, addressing any discrepancies or gaps in the patient data. Logical inference is then performed using a Prolog engine to generate diagnostic outcomes that strictly follow medical protocols. If the initial inference cycle does not yield definitive results, ProCDS iteratively updates the facts and rules, resolving mismatches and repeating the reasoning process. This neuro-symbolic integration enables ProCDS to offer transparent and precise clinical decision support. To fully evaluate ProCDS, we validated it in real-world OSAHS diagnostic scenarios and across open-source logical reasoning benchmarks to demonstrate its versatility. Our results show high accuracy and reliability. The key contributions of this work include:

- We propose a novel framework ProCDS, which adaptively refines rules and facts based on feedback from the Prolog engine to enhance both reasoning adaptability and accuracy.
- We implement ProCDS in a real clinical OSAHS diagnosis scenario, where it achieves superior diagnosis performance at the level of trained clinicians without performing task-specific prompt engineering.
- We also evaluate ProCDS across various open-source logical reasoning benchmarks, illustrating its adaptability and robustness across various reasoning domains.
- Our proposed ProCDS framework is constructed and evaluated based on the open-sourced LLMs for easy replication.

[4] The code is available at https://github.com/testlbin/procds.

2 Methods

2.1 Preliminary and Related Works

Instructions Following of Large Language Models There are several strategies to guide LLMs in generating specific outputs. In-context Learning (ICL) [24] uses example pairs $(\mathbf{x}_i, \mathbf{y}_i)$ to prompt pre-trained base models (p_θ), enabling few-shot learning for structured outputs. Techniques like supervised fine-tuning (SFT) [20] and reinforcement learning from human feedback (RLHF) [4] refine these models' ability to follow instructions by fine-tuning them on diverse tasks, resulting in an updated model $(p_{\bar{\theta}})$ that better adheres to instructions. The Chain-of-Thought (CoT) technique [22] enhances logical reasoning by guiding models through multi-step problems, improving accuracy and response detail. However, these methods don't always mitigate hallucination or ensure reasoning accuracy, especially in sensitive domains. Studies [14,31] report that reasoning steps $\mathbf{c}$ in CoT can be disorganized, failing to follow a structured logic sequence, which limits model performance. Additionally, missing information in few-shot demonstrations and inaccurate instructional data can worsen hallucinations in LLMs [11,13].

Neural-symbolic Framework with Prolog Logical Engine The integration of neural-symbolic frameworks has led to notable improvements in mathematical and reasoning tasks [12,27]. By incorporating logical engines like Prolog, LLMs can derive logical inferences that enhance output accuracy. Prolog [6] is a symbolic language system designed for rule-based reasoning within Horn Clause logic [3]. It uses a declarative programming paradigm, where computation logic is expressed through facts set $\mathcal{F}$ and rules set $\mathcal{R}$. Prolog's core reasoning relies on unification and backtracking. Unification matches predicate values by substituting variables, while backtracking allows Prolog to explore alternative solutions by revisiting previous steps. Typically, Prolog performs depth-first search to identify viable inference paths but can also enumerate all possible paths leading to the target results set $\mathcal{T}$. In medical decision support, such as OSAHS diagnosis, $\mathcal{T}$ can be clearly defined. However, variability in electronic health records (EHR) and diagnostic protocols complicates the creation of applicable rules $\mathcal{R}$ and facts $\mathcal{F}$, which are essential for building effective neural-symbolic systems for clinical support.

2.2 ProCDS

The proposed ProCDS contains two stages of processing to achieve the adaptive logical reasoning process using a logical engine with LLMs. Initially, the LLM extracts medical protocols and patient information, translating these into a structured format of rules and facts. This structured information is then enriched by aggregating additional related facts and rules, ensuring a comprehensive dataset for the logical engine. In the second stage, this information is transformed into executable code suitable for inference by the logical engine. Should the logical engine fail to provide valid results, the LLM iteratively refines and updates the information by generating supplement and revised rules and facts, thus filling in any gaps in the data for logical engine inference. The overview of ProCDS is shown in the Figure 1.

Stage 1: Rules and Facts Sets Extraction In the initial phase, we deploy an open-source LLMs, $p_{\bar{\theta}}$, characterized by parameters $\bar{\theta}$. Our goal is to facilitate rule-based diagnosis for OSAHS by first collating the medical guidelines $\mathbf{r}$ which include the gold-standard diagnostic criteria and associated rules for OSAHS. Subsequently, we collect the patient's EHR $\mathbf{x}$, which may vary across several dimensions and might be incomplete due to recording errors.

To begin, we design an initial extraction prompt, Prompt_{rf} , to systematically extract a potential set of rules $\mathcal{R}_I$ and facts $\mathcal{F}_I$ from $\mathbf{r}$ and $\mathbf{x}$ through:

$$\mathcal{R}_I, \mathcal{F}_I \sim p_{\bar{\theta}}[\cdot|\mathrm{Prompt}_{rf}(\mathbf{x}, \mathbf{r})]. \tag{1}$$

To enhance the adaptability of the system and ensure the sufficiency of reference rules and facts for subsequent logical inference, a second round of fact and rule augmentation is conducted. Using a modified prompt, Prompt_{r_rf}, we generate an additional set of rules $\mathcal{R}_S$ and facts $\mathcal{F}_S$ for supplementary purpose:

$$\mathcal{R}_S, \mathcal{F}_S \sim p_{\bar{\theta}}[\cdot|\mathrm{Prompt}_{r\_rf}(\mathbf{x}, \mathbf{r}, \mathcal{R}_I, \mathcal{F}_I)]. \tag{2}$$

Initially, we limit the process to a single iteration of fact and rule enhancement as delineated in Equation 2.

Stage 2: Iterative Neuro-Symbolic Integration In the second phase, the sets of rules $\mathcal{R}_S$ and facts $\mathcal{F}_S$ are translated into Prolog code, facilitating logical reasoning within the inference engine. Upon evaluating the performance of the model $p_{\bar{\theta}}$ used, it was demonstrated to effectively and accurately convert these rules and facts into executable Prolog code under a specifically designed prompt, Prompt_P:

$$\mathcal{R}_P, \mathcal{F}_P \sim p_{\bar{\theta}}[\cdot|\mathrm{Prompt}_P(\mathcal{R}_S, \mathcal{F}_S)], \tag{3}$$

where $\mathcal{R}_P$ and $\mathcal{F}_P$ denote the Prolog-compatible rules and facts. Given the inference target $\mathcal{T}$, which in this case involves diagnosing specific levels of OSAHS in clinical decision support scenarios, the Prolog engine generates reasoning outcomes:

$$\mathcal{O} = \mathrm{Prolog}(\mathcal{R}_P, \mathcal{F}_P, \mathcal{T}), \tag{4}$$

where $\mathcal{O}$ represents the set of results, including reasoning trajectories.

Considering the variability in patient information and the potential for incomplete EHR data, $\mathcal{R}_S$ and $\mathcal{F}_S$ may not always provide sufficient evidence for definitive Prolog inferences. Therefore, a dynamic update process for rules and facts is instituted whenever errors are identified in $\mathcal{O}$. This update mechanism is facilitated through another prompt, Prompt_U, which refines the generated rules and facts:

$$\mathcal{R}_U, \mathcal{F}_U \sim p_{\bar{\theta}}[\cdot|\mathrm{Prompt}_U(\mathcal{R}_P, \mathcal{F}_P, \mathcal{R}_S, \mathcal{F}_S)], \tag{5}$$

subsequently leading to a renewed logical reasoning cycle with the updated Prolog code:

$$\mathcal{O}_U = \text{Prolog}(\mathcal{R}_U, \mathcal{F}_U, \mathcal{T}). \tag{6}$$

This iterative refinement process can be continuously implemented until $\mathcal{O}_U$ yields error-free and reliable clinical decision support. However, in practical applications, even a single iterative update can substantially enhance the accuracy of the logical engine's inferences.

Model	**OSAHS**			**Proofwriter**	**GSM8K**
	Accuracy	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**
	No Mask	Mask	Keywords Mask		
BERT	92.18	86.46	81.11	-	-
PubMedBERT	89.94	77.22	79.72	-	-
ClinicalBERT	90.78	81.67	83.61	-	-
BioGPT	91.06	85.83	85.28	-	-
GatorTron	91.34	88.06	83.89	-	-
GPT-3.5-Turbo	85.18	79.18	82.30	30.92	41.11
with CoT	82.58	82.80	84.14	49.70	65.56
ProCDS (GPT3.5)	**99.55**	85.14	91.65	**78.55**	**70.74**
ProCDS (Llama3)	99.49	**91.43**	**95.32**	63.55	68.51

Table 1. Main experimental results ProCDS. The best results are marked in **bold**.

3 Experiment

3.1 Experimental Details

Experiment Setup We utilize Llama3-8B-Instruct [1] for ProCDS implementation. Inference is performed using the vLLM framework with Temperature=0.2, Top-p=1, and a maximum token limit of 2048 for controlled sampling. For logical symbolic reasoning, we employ the Prolog programming language with the SWI-Prolog inference engine [23]. We develop a symbolic solver using the pyswip[5] package to enable reliable batch-scale inference.

OSAHS Dataset Prior studies [17,21] demonstrate strong correlations between symptom severity and patient characteristics. We analyzed a de-identified dataset of 1,797 patients, approved by the Ethics Committee of the Eye & ENT Hospital of Fudan University (No.2022140). All data were anonymized and comply with privacy standards. Based on literature-derived indicators [15], we established 14 expert-validated rules to classify patients as high-risk (moderate/severe) or low-risk (normal/mild). Patients satisfying three or more rules are classified as high-risk.

General Datasets: To assess ProCDS general reasoning capability, we employed two benchmarks:

[5] https://github.com/yuce/pyswip

Model	Dataset	Accuracy (Before Verify)	Accuracy (After Verify)	Correction Rate
ProCDS (GPT3.5)	No Mask	98.99	99.55	83.33 (**12** → **2**)
	Mask	80.63	85.14	47.09 (**172** → **91**)
	Keywords Mask	89.81	91.65	73.33 (**45** → **12**)
	ProofWriter	73.20	78.55	53.26 (**199** → **93**)
	GSM8K	68.52	70.74	28.13 (**32** → **23**)
ProCDS (Llama3)	No Mask	99.38	99.49	25.00 (**8** → **6**)
	Mask	84.86	91.43	61.95 (**205** → **78**)
	Keywords Mask	92.93	95.32	93.75 (**48** → **3**)
	ProofWriter	56.45	63.55	29.66 (**372** → **230**)
	GSM8K	59.25	68.51	68.42 (**57** → **18**)

Table 2. After the Stage 2 error correction, the results of ProCDS. (A → B), marked in **bold**, indicates the correction from A errors to B errors ($CorrectionRate = error_{correct}/error_{total}$).

- **ProofWriter:** The ProofWriter dataset [19] challenges models with logical reasoning tasks, presenting premises and a hypothesis to verify in English. The task outcome can be classified as True, False, or Unknown, depending on the logical consistency with given facts and rules. The dataset is segmented into subsets by maximum proof depth. For our purposes, we engaged with the most demanding subset, allowing proof depths of ≤ 5, which comprises 482 rule sets and 2,000 randomly associated problems.
- **GSM8K:** The GSM8K dataset [5] serves as a benchmark for assessing and training models on mathematical reasoning skills. It features 8,000 problems suitable for grade-school level, spanning a diverse array of mathematical topics and problem types. We specifically utilized a subset of 270 annotated problems from this dataset [18] to evaluate the reasoning capability of ProCDS.

These benchmarks require multi-step logical and numerical reasoning capabilities that are directly analogous to real-world clinical decision-making processes, such as dosage calculations and rule-based differential diagnosis.

Baselines: To rigorously evaluate ProCDS efficacy, we employed various NLP models for patient risk classification. Note that models like BERT[7] are not designed for direct logical reasoning, making them unsuitable for ProofWriter and GSM-8K benchmarks. The baseline models are as follows:

- **BERT-based Models:** BERT[7]: Pre-trained transformer encoder fine-tuned for text classification tasks. PubMedBERT [9]: BERT variant trained from scratch on PubMed literature. ClinicalBERT[2]: BERT pre-trained on electronic health records for clinical text understanding. GatorTron[28]: Large-scale BERT-based model specialized in clinical data and medical terminology.

- **LLMs:** BioGPT[16]: Generative model trained on biomedical literature for domain-specific text generation. GPT-3.5-Turbo: Evaluated for direct inference, Chain-of-Thought reasoning, and framework integration.

Prompting Strategies: To assess the reasoning capabilities of LLMs under different prompt settings, we employed two prompting strategies widely used in LLM-based inference:

- **Direct Inference:** Generates outputs without additional reasoning constraints.
- **Chain-of-Thought[22]:** Enhances complex reasoning through structured, step-by-step logical prompting.

3.2 Experimental Results

To evaluate ProCDS robustness under real-world data incompleteness, we simulated three conditions:

- **No Mask**: Original complete data.
- **Random Mask 10%**: Randomly selecting 10% of patients and masking 2 information pieces for each selected patient to simulate missing information commonly observed in real-world EHRs. This setting tests the model's robustness against incomplete inputs.
- **Keywords Mask 10%**: Masking critical keywords (e.g., BMI, weight) that influence >3 diagnostic rules for 10% of patients. Keywords were identified by three clinicians through consensus following clinical guidelines.

As shown in Table 1, ProCDS achieves superior performance across all masking scenarios. In the No Mask condition, ProCDS attained 99.49% accuracy for OSAHS clinical decision support. Notably, GPT-3.5-Turbo's performance improved significantly when integrated with our framework, surpassing traditional methods like BERT and CoT. On ProofWriter and GSM8K datasets, ProCDS demonstrated comparable or superior performance to standalone GPT-3.5-Turbo, indicating effective generalization to common reasoning tasks.

Error Correction The verification process substantially enhanced model performance across all scenarios (Table 2). GPT-3.5-Turbo reduced error cases by an average of 57%, demonstrating strong analytical capabilities in logical reasoning tasks. ProCDS exhibited superior error correction in both clinical and mathematical reasoning domains.

4 Conclusion

In this paper, we introduce a novel neural-symbolic framework, ProCDS, designed to enhance clinical decision support using LLMs. Our framework addresses challenges related to missing information and the dynamic generation of facts and rules necessary for effective logical reasoning within a logical engine in Prolog. The ProCDS framework operates through a two-stage process by gathering possible rules and facts from provided information and iteratively refining the facts and rules based on the feedback from the logical engine. We have validated ProCDS using real-world scenarios, including the diagnosis of OSAHS, to verify its effectiveness in clinical decision support. Additionally, we tested ProCDS on open-source logical reasoning benchmarks to demonstrate its adaptability across various reasoning domains. The adaptability of ProCDS is particularly valuable for reasoning tasks requiring the processing of diverse and flexible natural language inputs from different individuals.

Acknowledgments. This work is supported by Shanghai Municipal Natural Science Foundation (23ZR1425400) and Shanghai Soft Science Project (25692114700).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. AI@Meta: Llama 3 model card (2024)
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78 (2019)
3. Chandra, A.K., Harel, D.: Horn clause queries and generalizations. The Journal of Logic Programming **2**(1), 1–15 (1985)
4. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
5. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
6. Colmerauer, A.: An introduction to prolog iii. Communications of the ACM **33**(7), 69–90 (1990)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 NAACL-HLT. pp. 4171–4186 (2019)
8. Goyal, S., Rastogi, E., Rajagopal, S.P., Yuan, D., Zhao, F., Chintagunta, J., Naik, G., Ward, J.: Healai: A healthcare llm for effective medical documentation. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining. p. 1167–1168. WSDM '24 (2024)
9. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing (2020)
10. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13418–13427 (June 2024)
11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12) (2023)

12. Lee, J., Hwang, W.: Symba: Symbolic backward chaining for multi-step natural language reasoning. arXiv preprint arXiv:2402.12806 (2024)
13. Li, Y., Wang, S., Ding, H., Chen, H.: Large language models in finance: A survey. In: Proceedings of the Fourth ACM International Conference on AI in Finance. p. 374–382. ICAIF '23 (2023)
14. Ling, Z., Fang, Y., Li, X., Huang, Z., Lee, M., Memisevic, R., Su, H.: Deductive verification of chain-of-thought reasoning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 36407–36433 (2023)
15. Liu, C., Chen, M.S., Yu, H.: The relationship between obstructive sleep apnea and obesity hypoventilation syndrome: a systematic review and meta-analysis. Oncotarget **8**(54), 93168 (2017)
16. Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.Y.: BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics **23**(6) (2022)
17. Ohdaira, F., Nakamura, K., Nakayama, H., Satoh, M., Ohdaira, T., Nakamata, M., Kohno, M., Iwashima, A., Onda, A., Kobayashi, Y., et al.: Demographic characteristics of 3,659 japanese patients with obstructive sleep apnea–hypopnea syndrome diagnosed by full polysomnography: associations with apnea–hypopnea index. Sleep and Breathing **11**, 93–101 (2007)
18. Ribeiro, D., Wang, S., Ma, X., Zhu, H., Dong, R., Kong, D., Burger, J., Ramos, A., Wang, W., Huang, Z., et al.: Street: A multi-task structured reasoning and explanation benchmark. arXiv preprint arXiv:2302.06729 (2023)
19. Tafjord, O., Dalvi, B., Clark, P.: ProofWriter: Generating implications, proofs, and abductive statements over natural language. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3621–3634 (2021)
20. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
21. Wang, Q., Zhang, C., Jia, P., Zhang, J., Feng, L., Wei, S., Luo, Y., Su, L., Zhao, C., Dong, H., et al.: The association between the phenotype of excessive daytime sleepiness and blood pressure in patients with obstructive sleep apnea-hypopnea syndrome. International journal of medical sciences **11**(7), 713 (2014)
22. Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 24824–24837 (2022)
23. Wielemaker, J., Schrijvers, T., Triska, M., Lager, T.: Swi-prolog. Theory and Practice of Logic Programming **12**(1-2), 67–96 (2012)
24. Wies, N., Levine, Y., Shashua, A.: The learnability of in-context learning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 36637–36651 (2023)
25. Wu, X., Li, Y.L., Sun, J., Lu, C.: Symbol-llm: Leverage language models for symbolic system in visual human activity reasoning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 29680–29691 (2023)
26. Xu, F., Wu, Z., Sun, Q., Ren, S., Yuan, F., Yuan, S., Lin, Q., Qiao, Y., Liu, J.: Symbol-llm: Towards foundational symbol-centric interface for large language models. arXiv preprint arXiv:2311.09278 (2023)
27. Yang, S., Li, X., Cui, L., Bing, L., Lam, W.: Neuro-symbolic integration brings causal and reliable reasoning proofs. arXiv preprint (2023)
28. Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., Zhang, Y., Magoc, T., Harle, C.A.,

Lipori, G., Mitchell, D.A., Hogan, W.R., Shenkman, E.A., Bian, J., Wu, Y.: A large language model for electronic health records. npj Digital Medicine **5**(1), 194 (2022)

29. Yang, Z., Xu, X., Yao, B., Rogers, E., Zhang, S., Intille, S., Shara, N., Gao, G.G., Wang, D.: Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **8**(2) (may 2024)
30. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al.: Siren's song in the ai ocean: a survey on hallucination in large language models. arXiv preprint arXiv:2309.01219 (2023)
31. Zheng, H.S., Mishra, S., Chen, X., Cheng, H.T., Chi, E.H., Le, Q.V., Zhou, D.: Take a step back: Evoking reasoning via abstraction in large language models. In: The Twelfth International Conference on Learning Representations (2024)