Extracting problem-solving knowledge from LLMs with reasoning abilities

Maxime Haurel[1,2]

[1]Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France
[2]Datanello, Nancy, France

Abstract

Automatic Knowledge Acquisition (AKA) aims to automate the process between domain experts and knowledge engineers, that is to create a domain Knowledge Base (KB). Such automation is necessitated because the collaboration between domain experts and knowledge engineers is costly. Modern approaches use Large Language Models (LLMs) simulating domain experts to create ontologies and knowledge graphs. Recently, LLMs with reasoning abilities received attention due to their great performances on several benchmarks. These LLMs output reasoning traces that lead to the answer. This early stage PhD thesis (started 7 months ago) focuses on the use of those reasoning traces in the automatic construction of a Knowledge Base (KB), under the assumption that they express the knowledge necessary to solve the problem prompted to the LLM. To achieve this, a first step consists in obtaining from the LLM reasoning traces that are expressed in a well-defined formalism. The results of our initial experiments show that, while some models are able to generate reasoning traces backed by formally expressed knowledge, there is still room for improvement. The remainder of this PhD will therefore involve improving LLMs with reasoning abilities to increase their capability to express the knowledge they used for problem solving in a way that is exploitable for AKA, coherent, and valid.

Keywords

automatic knowledge acquisition, large language models, reasoning models, prolog

CEUR-WS.org/Vol-4085/paper20.pdf

1. Problem statement

Automatic Knowledge Acquisition (AKA) is a task known for being time-consuming. Indeed, domain experts and knowledge engineers need to collaborate to extract domain knowledge and represent it in a Knowledge Base (KB) in a structured way [1]. The literature has recently shown that building KBs can be supported by Large Language Models (LLMs), both semi-automatically [2] and automatically [3, 4]. In studies using LLMs for AKA, the majority of approaches use LLMs without reasoning abilities [2, 3]. Indeed, LLMs with reasoning abilities (or reasoning models) only appeared recently and very few works have yet exploited them for KB construction [4]. In this work, we look specifically into using the reasoning abilities of LLMs in AKA, that is to extract knowledge from those LLMs, and more specifically problem-soving knowledge. To clarify what we mean by reasoning abilities in LLMs, they are trained with Reinforcement Learning (RL) [5] which enables them to output their reasoning process as natural language in a `<think></think>` section. They have received particular attention since the publication of the DeepSeek-R1 model in January 2025 [6] and show remarkable results on numerous benchmarks. The RL training of the LLM has been designed to reward results where the reasoning section (`think`) includes text representing a "chain of thought" leading to the answer. An example is shown in Table 1.

In this PhD Thesis, we make the hypothesis that this textual and non-formal reasoning can be analyzed to extract problem-solving domain knowledge on which the LLM relies to construct its answer. Unsurprisingly, considering how recently LLMs with reasoning abilities came out, they have so far rarely been used. This work aims to explore the viability of automatically creating reliable KBs by leveraging such LLMs. The ultimate goal is to make an LLM with reasoning abilities able to automatically manage a KB's lifecycle (refine, reuse, reason over). Achieving this goal will contribute towards the development of both AKA (i.e. facilitating the construction of bigger, more reliable KBs) and LLMs (i.e. making LLMs

---

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan

maxime.haurel@loria.fr (M. Haurel)
0009-0006-8010-808X (M. Haurel)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

more interpretable). In the next sections, we outline several approaches to verify our hypotheses. In the end of this document, we enumerate the next steps of this work.

> **<think>** Okay, so I'm trying to figure out the possible heart condition based on the symptoms described. The user mentioned sudden weakness in the left arm and leg, which makes me think of a stroke or possibly a TIA (transient ischemic attack). But then there's also recent long-distance travel and a swollen, tender right lower leg.
> First, the left-sided weakness points towards neurological issues, likely related to blood flow. A stroke or TIA is a common cause for sudden weakness in one side of the body. However, the user also provided other symptoms that I need to consider together.
> The long-distance travel makes me think about prolonged sitting or physical activity during the trip. This could contribute to deep vein thrombosis (DVT), especially if there's a history of blood clots or risk factors like obesity, age over 40, or being sedentary for long periods.
> Now, the right lower leg symptoms—swollen and tender—are classic signs of DVT. DVT occurs when blood clots form in the legs and can travel to the lungs (causing a pulmonary embolism) or other organs. If someone experiences sudden weakness on one side along with DVT, it might be related to an embolism blocking a blood vessel in the brain, leading to a stroke.
> So, putting this together, the person might have experienced a pulmonary embolism from DVT, which then caused a stroke-like syndrome. This would explain both sets of symptoms: the left-sided weakness from the stroke and the right leg issues from DVT. **</think>**
> <answer> The most likely cardiac abnormality is a pulmonary embolism secondary to deep vein thrombosis (DVT). The recent long-distance travel increases the risk of DVT, which can lead to a pulmonary embolism. This condition can cause sudden neurological symptoms like left-sided weakness due to an embolism blocking a cerebral artery, mimicking a stroke or TIA.</answer>

Table 1
An example of the raw output obtained from the 8B distilled version of DeepSeek-R1. It explicitly shows the reasoning content (`<think>` tags in blue). The input problem is: *Given the symptoms of sudden weakness in the left arm and leg, recent long-distance travel, and the presence of swollen and tender right lower leg, what specific cardiac abnormality is most likely to be found upon further evaluation that could explain these findings?*

2. Importance

AKA is a crucial question in the artificial intelligence community focusing on knowledge-based systems. Indeed, representing unstructured domain knowledge into structured information is a process that generates both human and time costs. This bottleneck limits our ability to build KBs that are both large-scale [7] and reliable [7]. We make the hypothesis that using LLMs to automatically acquire knowledge and manage a KB's lifecycle will reduce this issue.

Our approach aims to automatically build a KB, using the reasoning traces produced by LLMs with reasoning abilities as the expression of knowledge. We want to obtain a formalized version of the reasoning traces in the objective of building a KB with a formalism that has proven to be efficient for this task in the knowledge representation and reasoning community. In addition, LLMs with reasoning abilities have recently emerged, in particular with DeepSeek-R1 in January 2025 [6]. While these models are able to show a form of explanation compared to LLMs without reasoning abilities, LLMs in general (i.e. with or without reasoning abilities) are still operating as black boxes [8, 9, 10]. Our approach will help us better understand these models, contributing to interpretability. Indeed, our approach could explain a prediction through logical reasoning, or even validate/invalidate the results obtained based on the coherence of the logical reasoning.

Our research is motivated by a problem encountered in the HR domain. The issue is to identify indicators of soft skills that are expressed in a text that summarizes the decisions made by an individual in the role of a manager facing a challenging situation (e.g. dealing with an angry customer due to a problem of shipping). This task is highly knowledge-intensive as it requires solid psychology-related knowledge from an expert. Also, no labeled data is available for this task of predicting soft skills indicators from a text. Moreover, we face the issue of the availability of domain experts qualified on the subject. This is why we require a method that is both automatic and explainable. In addition, the HR domain is a high-stakes domain as listed in the EU AI Act [11]. Specifically, the document states that: "*AI systems intended to be used to make decisions affecting terms of work-related relationships, the promotion or termination of work-related contractual relationships, to allocate tasks based on individual behaviour or personal traits or characteristics or to monitor and evaluate the performance and behaviour of persons in such relationships*" are classified as high risk. We then need a method that provides answers given a complex problem (e.g. predicting soft skills indicators from a text) using expert knowledge formally defined in a KB while providing explanations regarding the reasoning process used by the model to reach the answer. Thus, building a problem-solving knowledge focused KB automatically with an LLM to then use this KB to reason over complex problems is promising. While our objective is to use LLMs with reasoning abilities as domain experts to manage a KB's lifecycle, we are aware of the ethical considerations [12] and we will deal with these considerations when developing the framework. While other approaches exist, our approach is intended to reduce the barriers of the lack of domain experts and the lack of annotated data. In the study described in this paper, we take examples from the medical domain from a public dataset. This kind of knowledge requires solid reasoning to provide reliable answers [13], making it close to the HR domain. Furthermore, this dataset has proven to be a good support to exert reasoning from LLMs [13].

3. Related work

Our work is placed in AKA, but with a set of constraints. As said earlier, our objective is to capture knowledge useful in the resolution of a specific (naturally knowledge-intensive) problem, such as the prediction of soft skill indicators. Additionally, the knowledge that we need is not stored in any document. We do not have access to large quantities of past cases of such problems solved, and the solution is not automatically verifiable. Our interest in LLMs therefore comes from the assumption that, as an effect of their training, they have captured both common sense and domain-specific knowledge, which means that they could, at least partly, take the role of a domain expert. Explainability is a crucial consideration, given the high level of risk stated by the EU AI act in the HR domain [11]. Relying on explicitly formalized knowledge in a KB is therefore favored here. Hence, here, we review existing work in AKA, especially with LLMs. We also investigate the literature to review the formalisms used by the knowledge representation community.

Automatic Knowledge Acquisition Several works have focused on building KBs automatically. We identified two approaches in the AKA community: document (i.e. natural language) processing approach and learning (i.e. statistical) approach. As an example in document processing, [14] proposes to automatically construct a KB through document parsing, which therefore requires access to a corpus of relevant documents. As an example in a learning approach, [15] compares an inductive learning system to a naive bayesian learning system. The former generates knowledge as a decision tree, while the latter generates a table of conditional and prior probabilities. Both systems (inductive learning and naive bayesian learning) are based on a collection of concrete problems, solved by experts in the past. Here too, the applicability of the method is limited to cases where a base of previously solved problems is available, which is not the scenario on which we focus.

Automatic Knowledge Acquisition with LLMs LLMs are trained on a vast corpus of data, making them able to leverage a great amount of cross-domain knowledge. They even surpass humans in some fields such as coding [16], generating stories [17], and answering genetic-related questions [18].

The following related works rely on the idea that LLMs can be exploited as a source of knowledge, similarly to domain experts. Recently, [2] proposed to use LLMs to semi-automatically (i.e. with a human involved in the loop) construct an ontology and a Knowledge Graph (KG). Their approach relies on the generation of competency questions (CQs) to create an ontology and the answering of CQs to construct the KG. Other works have used LLMs to automatically construct a KG [3], where LLMs play the role of domain experts. Their solution is a framework made up of 3 levels: *Generator*, *Verifier*, and *Pruner*. For a specified entity (e.g. a domain name), *Generator* retrieves the most relevant triples from an open-source encyclopedic KG. Then, *Verifier* identifies and filters out the erroneous triples generated by the LLM at the previous level (i.e. Generator). Finally, *Pruner*, a binary classifier, predicts for each tail entity (object) from each correct triple, whether the entity should proceed to generate the next-level KG (i.e. continue the loop) or cease the generation. This work shows that LLMs can produce better KGs than the previous state-of-the-art methods and makes a great step towards AKA as no human is involved. Thus, it supports our research directions. However, as expressed above, our objective concerns the acquisition of knowledge that is useful for the resolution of specific problems, while most of the previous work on using LLMs for AKA has focused on ontologies and KGs, which are more general and abstract.

Automatic Knowledge Acquisition with reasoning abilities in LLMs Recent approaches utilize RL to influence the behavior of an LLM. The Group Relative Policy Optimization (GRPO) algorithm [5] provides an efficient way of training the LLM to maximize a reward with reinforcement learning. This algorithm is the basis behind the DeepSeek-R1 model [6] that achieved great scores on several benchmarks and opened the way to LLMs with reasoning abilities. The GRPO algorithm plays a key role as it maximizes the reward designed by the DeepSeek team. This reward aims to influence the behavior of the LLM so that the reasoning is outputted within `<think></think>` tags. Released in April 2025, the Cogito LLM [19] has received attention thanks to great performance on benchmarks. Since such models with reasoning abilities are quite recent, few works in the literature used them for AKA. [4, 20] show the use of LLMs with reasoning abilities to generate ontologies. They represent the knowledge in the OWL language after giving CQs as input to the LLM.

Knowledge Representation and Reasoning While many knowledge representation formalisms exist, Prolog is a good candidate for modeling symbolic representations of reasoning problems [21, 22, 23, 24, 25, 26, 27]. For example, [26] improves the logical reasoning of LLMs through a framework notably including a *Translator* module that aims to parse given premises and a question statement with an LLM into a symbolic format adopting Prolog's grammar. Another approach [27] uses Vadalog [28], a language derived from Datalog [29]. In the study, the Vadalog language is used to describe the knowledge graph and act as an intermediate layer between the natural language question formulated by the user and the actual knowledge represented as a graph.

In the scientific literature, Prolog has been widely used. For the application we are tackling throughout (i.e. human resources) this thesis, we believe that Prolog is the best candidate to represent knowledge for solving problems. In addition, Prolog is a logical formalism that can be expressed as text, which is crucial for asking an LLM to output it. Also, Prolog being widely represented among logical formalisms, we think that an LRM will be able to output it properly. While Prolog has many advantages, a limitation is that representing probabilities is not possible. To overcome this limitation and if we need a probabilistic representation of knowledge, we will consider using ProbLog [30], an extension of Prolog. For a Prolog clause, ProbLog allows to add probabilistic information. Finally, we plan to experiment with other types of formalisms than Prolog in the context of AKA as future work.

4. Research question(s) and hypotheses

This thesis aims to contribute to AKA by representing the knowledge in KBs by means of LLMs with reasoning abilities. More precisely, we aim to answer the following question: *How to leverage knowledge expressed in the reasoning traces of a LLM with reasoning abilities to automatically create and manage a domain KB?*

Developing the research question, we postulate the hypotheses described below.

H1: Given a defined reasoning formalism, LLMs with reasoning abilities are able, possibly with additional training, to generate a reasoning trace in this formalism.

LLMs have been used to generate Prolog outputs [25] but no works on this task have been published using LLMs with reasoning abilities, probably due to their recency. In a first step, we aim to assess whether such LLMs can generate syntactically valid Prolog in the reasoning part. The associated experiment is detailed in Section 5.

H2: Reinforcing the ability of the LLM to generate formalized knowledge for reasoning traces about a problem strengthens its reliability to generate valid answers, and valid knowledge.

This hypothesis necessitates an interest in both the reasoning traces for a given problem and, more generally, the KB that is produced by solving multiple problems. Regarding the reasoning traces, we start by introducing a measure s that allows us to measure how much a code is syntactically valid in a given language. We initially applied it to the Prolog language with the idea that further work will need to be done to also assess the semantic validity of the generated Prolog. Assessing the quality of a KB is much more complicated and will be the focus of further work. This kind of evaluation already exists and it involves several steps [1]:

- Debugging : detecting inconsistencies (conflict, redundancy, subsumption)
- Checking for completeness : detecting incompleteness (missing rules)

In the next section (Section 5), we begin by working on the first hypothesis. We designed an experiment to assess if each LLM tested is able to generate the reasoning trace in the syntax of the Prolog formalism. To achieve that, we use the measure s and, in the rest of the PhD, we will introduce other measures to evaluate the quality of the created KB.

5. Preliminary results

Among the available (i.e. open-weight) LLMs with reasoning abilities, we consider DeepSeek-R1, and Cogito, both in their distilled versions due to their computational costs. Concretely, we test the 1.5B[1] and 8B[2] distilled versions of DeepSeek-R1 and the 3B[3] and 8B[4] distilled versions of Cogito. We did not consider QwQ [31] or any larger model as they are too large for the resources to which we had access at this point. As we aim to fine-tune the model in future work, closed-source LLMs are not considered here.

Prompting A problem (i.e. a question whose answer requires reasoning) is prompted to one of the four LLMs, instructing it to express its reasoning in Prolog. In these preliminary experiments, this problem is taken from the medical-o1-dataset-SFT [13]. The problem is inserted in a prompt inspired by the one used to train DeepSeek [6], as shown in Table 3. We evaluate the quality of the Prolog generated in the reasoning part using the reward score s defined below. We repeat the process 10 times for each LLM for the same problem to get an averaged overview of the quality of the Prolog generated in the reasoning.

Given a parser[5], s is defined as

$$s = \begin{cases} 1 - (m/n) & \text{if } n > 0 \\ 1 & \text{else} \end{cases}$$

where n is the number of errors detected by the parser and m is the number of lines of the reasoning traces. s is equal to 0 if the LLM did not return anything in the reasoning.

Table 2 showcases the results obtained. The low s values obtained are justified by the fact that LLMs are not specifically trained for this task. However, we did expect to see better results from the DeepSeek-R1 models. When investigating these results, the reasoning is always blank for DeepSeek-R1:8B. Actually, no matter the version of DeepSeek-R1, and despite the number of tries with DeepSeek-R1, we did not achieve to incentivize the model to generate Prolog in the reasoning. We notice that the Cogito models

---

[1] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B

[2] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B

[3] https://huggingface.co/deepcogito/cogito-v1-preview-llama-3B

[4] https://huggingface.co/deepcogito/cogito-v1-preview-llama-8B

[5] Here we use https://github.com/akvnn/prolog-parser as the base of the parser.

are performing great on our measure, Cogito:3B having better performances than Cogito:8B on average. Even though the measure s is low, it is a measure evaluating the syntax of the Prolog code, it does not evaluate how much the Prolog code is meaningful. The meaningfulness of the Prolog code generated will be assessed when we evaluate the KB in the future of this thesis.

Model	Mean	Std	Min	Max
cogito:3b	0.3621	0.3619	0.0500	0.8333
cogito:8b	0.2406	0.3115	0.0500	0.7727
deepseek-r1:1.5b	0.0500	0.0000	0.0500	0.0500
deepseek-r1:8b	0.0000	0.0000	0.0000	0.0000

Table 2
Descriptive statistics of the measure s per model.

> A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<think>` `</think>`, i.e., `<think>` reasoning process here `</think>` tags. All reasoning inside `<think>` must be done in Prolog style: define facts, rules, and queries like a Prolog program. Use step-by-step logic and inference. Do not explain in natural language — only use Prolog. User: prompt. Assistant:

Table 3
Prompt template for Prolog reasoning. In blue are the modifications we made to the original template. In red is the place of the problem that the LLM need to reason about.

LLM Fine-Tuning The results obtained in the prompting phase encourage us to pursue with fine-tuning the Cogito LLM. The Prolog code generated in the reasoning of Cogito compared to DeepSeek-R1 makes us think that there is room for improvement. We plan to use the GRPO algorithm [5] (i.e. the same algorithm behind DeepSeek-R1) to fine-tune Cogito, using the score s as the reward to maximize. This fine-tuning is the next step in the thesis.

6. Evaluation

To answer our research question, we will focus on the same dataset (i.e. the medical-o1-reasoning-SFT dataset). However, to go beyond the syntactic evaluation we realized in Section 5, a first step will be to build a first KB by making an LLM reason over more than one problem. Taking inspiration from the literature [1], we will then introduce measures to validate the automatically acquired knowledge. In a second step, we plan to test our method on the use case in the HR domain. To validate the automatically acquired knowledge, we will survey practitioners of the domain.

7. Reflection and future work

As mentioned in Section 5, we plan to fine-tune the Cogito LLM with the GRPO algorithm, using the measure s as the reward to train the LLM to more robustly generate syntactically valid Prolog in its reasoning traces. One objective of our initial experiments was to check whether existing models had some level of ability on which such fine-tuning could build to achieve this task. As we have shown, this appears to be true for Cogito, but not for DeepSeek-R1. The next experiment to be carried out will therefore have for objective to test how much the ability of Cogito to generate reasoning traces as Prolog can be improved through fine-tuning. This will be the first step into an automatic management of a KB's lifecycle. We plan to build on this capability to extract domain knowledge, dealing with semantic validity and coherence over multiple problems in addition to syntactic validity. In addition, as mentioned earlier in this paper, we chose to use Prolog as a first formalism with which to test LLMs. An expected contribution will also include understanding which other formalisms might be more suitable for expressing the reasoning traces of LLMs. We also anticipate the need for a larger, more robust framework as showcased in related studies [3, 26, 27]. Building a larger framework will allow us to evaluate all steps independently in order to achieve the automatic construction of a robust KB. This framework, being applied to the HR domain, will also need a formalization, ensuring a high robustness.

As mentioned above, we need to go beyond evaluating the syntactic validity of the generated reasoning trace, and also address the quality of the KB created from combining such knowledge used for multiple, related problems. Hence, a bibliographic study on the evaluation of the quality of KBs is one of our priorities. In the long-term view of this research, we will use the formalism generation capability of the LLM in the reasoning to validate or invalidate its final prediction. This will act as a safeguard to flag invalid and/or illogical answers. We therefore anticipate for our method to also contribute to interpretability, providing a verifiable link between the reasoning trace provided and the generated answer. Indeed, while natural language in the reasoning, as it is currently expressed, provides a way of better understanding whether the answer might or might not be correct, our approach could provide a way to automatize this verification.

Acknowledgments

The author acknowledges Pr. Armelle Brun, professor of computer science at Université de Lorraine, Pr. Mathieu d'Aquin, professor of computer science at Université de Lorraine as supervisors for this PhD, and Dr. Yacine Abboud, director at Datanello. This research is funded by an ANRT grant through the CIFRE system with Datanello as the industrial collaborator.

Declaration on Generative AI

During the preparation of this work, the author used GPT-4 to generate LaTeX formulas and Writefull to review the syntax of the document. After using these tools, the author reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Suwa, A. C. Scott, E. H. Shortliffe, An approach to verifying completeness and consistency in a rule-based expert system, Ai Magazine 3 (1982) 16–16. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/377.

[2] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An LLM supported approach to ontology and knowledge graph construction, 2024. URL: http://arxiv.org/abs/2403.08345. doi:`10.48550/arXiv.2403.08345`, arXiv:2403.08345 [cs].

[3] H. Chen, X. Shen, Q. Lv, J. Wang, X. Ni, J. Ye, SAC-KG: Exploiting Large Language Models as Skilled Automatic Constructors for Domain Knowledge Graphs, 2024. URL: http://arxiv.org/abs/2410.02811. doi:`10.48550/arXiv.2410.02811`, arXiv:2410.02811 [cs].

[4] A. S. Lippolis, M. J. Saeedizade, R. Keskisärkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, A. G. Nuzzolese, Ontology Generation using Large Language Models, 2025. URL: http://arxiv.org/abs/2503.05388. doi:`10.48550/arXiv.2503.05388`, arXiv:2503.05388 [cs].

[5] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, 2024. URL: http://arxiv.org/abs/2402.03300. doi:`10.48550/arXiv.2402.03300`, arXiv:2402.03300 [cs].

[6] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong,

K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, Z. Zhang, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL: http://arxiv.org/abs/2501.12948. doi:10.48550/arXiv.2501.12948, arXiv:2501.12948 [cs].

[7] Q. He, J. Yu, W. Wang, Large Language Model-Enhanced Symbolic Reasoning for Knowledge Base Completion, 2025. URL: http://arxiv.org/abs/2501.01246. doi:10.48550/arXiv.2501.01246, arXiv:2501.01246 [cs].

[8] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, M. Du, Explainability for Large Language Models: A Survey, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–38. URL: https://dl.acm.org/doi/10.1145/3639372. doi:10.1145/3639372.

[9] C. Singh, J. P. Inala, M. Galley, R. Caruana, J. Gao, Rethinking Interpretability in the Era of Large Language Models, 2024. URL: http://arxiv.org/abs/2402.01761. doi:10.48550/arXiv.2402.01761, arXiv:2402.01761 [cs].

[10] F. Xu, Q. Hao, Z. Zong, J. Wang, Y. Zhang, J. Wang, X. Lan, J. Gong, T. Ouyang, F. Meng, C. Shao, Y. Yan, Q. Yang, Y. Song, S. Ren, X. Hu, Y. Li, J. Feng, C. Gao, Y. Li, Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models, 2025. URL: http://arxiv.org/abs/2501.09686. doi:10.48550/arXiv.2501.09686, arXiv:2501.09686 [cs].

[11] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance. (????).

[12] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, Y. Zhang, Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity, 2023. URL: http://arxiv.org/abs/2310.07521. doi:10.48550/arXiv.2310.07521, arXiv:2310.07521 [cs].

[13] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, B. Wang, HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs, 2024. URL: http://arxiv.org/abs/2412.18925. doi:10.48550/arXiv.2412.18925, arXiv:2412.18925 [cs].

[14] Y. Y. Tang, C. D. Yan, C. Suen, Document processing for automatic knowledge acquisition, IEEE Transactions on Knowledge and Data Engineering 6 (1994) 3–21. URL: https://ieeexplore.ieee.org/abstract/document/273022. doi:10.1109/69.273022.

[15] I. Kononenko, Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition, Current trends in knowledge acquisition 8 (1990) 190. URL: https://books.google.fr/books?hl=fr&lr=&id=5GgR1bJITVUC&oi=fnd&pg=PA190&dq=automatic+knowledge+acquisition&ots=uPIhpo4O2u&sig=jgrmxraRnFvJ0aqYPMRZb8srZ2E, publisher: IOS Press Amsterdam.

[16] W. Hou, Z. Ji, Comparing Large Language Models and Human Programmers for Generating Programming Code, Advanced Science 12 (2025) 2412279. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202412279. doi:10.1002/advs.202412279, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/advs.202412279.

[17] Z. Zhao, S. Song, B. Duah, J. Macbeth, S. Carter, M. P. Van, N. S. Bravo, M. Klenk, K. Sick,

A. L. S. Filipowicz, More human than human: LLM-generated narratives outperform human-LLM interleaved narratives, in: Proceedings of the 15th Conference on Creativity and Cognition, C&amp;C '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 368–370. URL: https://dl.acm.org/doi/10.1145/3591196.3596612. doi:10.1145/3591196.3596612.

[18] D. Duong, B. D. Solomon, Analysis of large-language model versus human performance for genetics questions, European Journal of Human Genetics 32 (2024) 466–468. URL: https://www.nature.com/articles/s41431-023-01396-8. doi:10.1038/s41431-023-01396-8.

[19] DeepCogito, Cogito v1 Preview Introducing IDA as a path to general superintelligence, 2025. URL: https://www.deepcogito.com/research/cogito-v1-preview.

[20] A. S. Lippolis, M. J. Saeedizade, R. Keskisarkka, A. Gangemi, E. Blomqvist, A. G. Nuzzolese, Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation, 2025. URL: http://arxiv.org/abs/2504.17402. doi:10.48550/arXiv.2504.17402, arXiv:2504.17402 [cs].

[21] L. Pan, A. Albalak, X. Wang, W. Y. Wang, Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning, 2023. URL: http://arxiv.org/abs/2305.12295. doi:10.48550/arXiv.2305.12295, arXiv:2305.12295 [cs].

[22] S. Yang, X. Li, L. Cui, L. Bing, W. Lam, Neuro-Symbolic Integration Brings Causal and Reliable Reasoning Proofs, 2025. URL: http://arxiv.org/abs/2311.09802. doi:10.48550/arXiv.2311.09802, arXiv:2311.09802 [cs].

[23] W. F. Clocksin, C. S. Mellish, Programming in Prolog, Springer Science & Business Media, 2003. Google-Books-ID: VjHk2Cjrti8C.

[24] P. Körner, M. Leuschel, J. Barbosa, V. S. Costa, V. Dahl, M. V. Hermenegildo, J. F. Morales, J. Wielemaker, D. Diaz, S. Abreu, G. Ciatto, Fifty Years of Prolog and Beyond, Theory and Practice of Logic Programming 22 (2022) 776–858. URL: https://www.cambridge.org/core/journals/theory-and-practice-of-logic-programming/article/fifty-years-of-prolog-and-beyond/3A5329B6E3639879301A6D44346FD1DD. doi:10.1017/S1471068422000102.

[25] X. Yang, B. Chen, Y.-C. Tam, Arithmetic Reasoning with LLM: Prolog Generation & Permutation, 2024. URL: http://arxiv.org/abs/2405.17893. doi:10.48550/arXiv.2405.17893, arXiv:2405.17893 [cs].

[26] J. Xu, H. Fei, M. Luo, Q. Liu, L. Pan, W. Y. Wang, P. Nakov, M.-L. Lee, W. Hsu, Aristotle: Mastering Logical Reasoning with A Logic-Complete Decompose-Search-Resolve Framework, 2024. URL: http://arxiv.org/abs/2412.16953. doi:10.48550/arXiv.2412.16953, arXiv:2412.16953 [cs].

[27] T. Baldazzi, L. Bellomarini, S. Ceri, A. Colombo, A. Gentili, E. Sallinger, "Please, Vadalog, tell me why": Interactive Explanation of Datalog-based Reasoning, volume #PLACEHOLDER_PARENT_METADATA_VALUE#, OpenProceedings, 2024, pp. 834–837. URL: https://repositum.tuwien.at/handle/20.500.12708/210849. doi:10.48786/edbt.2024.82, accepted: 2025-02-04T16:48:04Z.

[28] L. Bellomarini, D. Benedetto, G. Gottlob, E. Sallinger, Vadalog: A modern architecture for automated reasoning with large knowledge graphs, Information Systems 105 (2022) 101528. URL: https://www.sciencedirect.com/science/article/pii/S0306437920300351. doi:10.1016/j.is.2020.101528.

[29] S. Abiteboul, R. Hull, V. Vianu, Foundations of Databases, Addison-Wesley, 1995.

[30] L. D. Raedt, A. Kimmig, H. Toivonen, ProbLog: A Probabilistic Prolog and its Application in Link Discovery (2007).

[31] Q. Team, Qwq-32b: Embracing the power of reinforcement learning, URL: https://qwenlm. github. io/blog/qwq-32b (2025).