New directions to be explored in integrating Large Languge Models for knowledge elicitation

Vlad Luca Sebastian Spataru

Supervisor(s): Ujwal Gadiraju, Shreyan Biswas

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Vlad Luca Sebastian Spataru
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Shreyan Biswas, Ricardo Marroquim

Abstract

As many entities aim to participate in the ongoing AI race to gain competitive advantages, there is a risk of creating knowledge gaps by overlooking fundamental steps in the research and development processes. This paper aims to bridge the knowledge gap in the domain of Large Language Model (LLM) integrations for knowledge elicitation by performing a systematic literature review using the PRISMA workflow. Through an analysis of 17 research papers, this study identifies new directions, including tools for education, knowledge curation, and factual information. Additionally, the research highlights key benefits and concerns associated with each direction, providing further understanding of the potential and challenges of LLM integrations for knowledge elicitation.

1 Introduction

With the recency of the topic of Large Language Models (LLMs) and Artificial Intelligence (AI) many entities are looking to gain a competitive advantage in the field, and awareness must be raised of the risk of these entities cutting corners when developing new technology [5]. This can potentially lead to companies prioritizing gaining a competitive advantage over following standard research and development processes. Moreover, skipping fundamental steps in research and development may leave knowledge gaps behind that can hardly be filled in without in-depth analysis and systematic approaches. Within the multiple knowledge gaps caused by this race, we identify a lack of research papers synthesizing new directions that can be followed for further exploration and integration.

The topic of LLMs and AI is vast, and many different domains can be explored. The direction that will be the focus of this paper is the integration of LLMs for knowledge elicitation. When coupling the term knowledge elicitation and Large Language Models several relevant applications can be identified such as chatbots with one of the popular integrations being ChatGPT and Games with a purpose (GWAPs). Therefore, this paper aims to identify and categorise new directions for integrating Large Language Models for knowledge elicitation and provide an overview of what the industry and research community should potentially look further into when exploring this particular domain.

Research Question

Considering the previously mentioned knowledge gap on this topic and the motivation to bridge it, this paper intends to contribute to the research community by synthesizing possible new directions for integrations of LLMs for knowledge elicitation. Identifying new directions aims to drive the research community and industry into developing effective applications for end-users that can potentially have a positive impact on society. Therefore, this research attempts to answer the following research question:

> *What new directions can be explored in integrating Large Language Models for knowledge elicitation?*

Furthermore, sub-questions have been identified to define and clarify further the objectives in analyzing and finding directions. As this research will be a literature review, we will narrow down the new directions that this research aims to identify to concepts and ideas presented in available research but require further exploration or real-world integration. Thus, the first sub-question clarifies the domain from which the new directions will be extracted.

> *What directions has the research community and industry identified that have yet to be thoroughly researched or productized?*

Furthermore, to have a basis for a closer analysis and discussion on the directions identified, sub-questions have been derived to look into the benefits and concerns of each topic.

> *What are the possible strengths and benefits of exploring and integrating the identified directions?*
>
> *What possible concerns exist with the integration of the identified directions?*

Therefore, we clarify the scope of identifying, analyzing, and objectively criticizing what literature presents as the benefits and concerns of the possible new directions.

Structure

In the following section, a brief introduction to related previous work is given. Then, it will outline the methodology utilised for answering the outstanding research question. The fourth chapter will present the background concepts that are at the foundation of the research, followed by the main findings and analysis of the relevant literature. Next implications, reflection, and future work are presented in the discussion section, followed by the limitations of the study. Lastly, the paper summarises the key findings and contributions in the conclusion.

2 Background & Related work

Knowledge elicitation contains multiple techniques and approaches that aim to extract knowledge from a domain expert [25]. Also referred to as knowledge extraction within the research domain, it is tightly related to similar concepts such as knowledge acquisition and knowledge engineering, each being a sub-process of the other [6]. Large Language Models are systems that can solve a variety of different tasks comparable to human-level performance that can process and generate text with comprehensible communication [17].

These two key concepts, knowledge elicitation, and large language models, sit at the foundation of this paper. By looking at the description of the two, one can identify a variety of applications that have been researched, developed, and are used even by average consumers who might not be familiarised with the domain itself.

One of the most popular examples is ChatGPT, a versatile application of the GPT LLMs (GPT-3.5, GPT-4, or GPT-4o) that aims to perform a variety of different tasks such as text generation, analysis, and translation [23]. This application has become one of the most popular integrations worldwide due to its ease of use for non-experts [23].

Games with a purpose are also relevant as an example despite the limited set of applications available currently for consumers. GWAPs are games created with the scope of solving large-scale problems that computers cannot solve on their own [29]. Previous research work in the domain categorises Games With a Purpose (GWAPs) as a good example of an integration with LLMs, indicating that they are efficient in eliciting diverse knowledge, with some games being configurable to feed downstream AI applications to make use of it [4]. One example application that research has been performed on is FindItOut. It is a game that aims to extract and process data used by players, in such a way that it can be exploited by AI systems [24]. Furthermore, other integrations in video games have also been explored as proof of concepts and tools for the advancement of AI research through the use of LLM-based game agents trained on extracted data from gameplay of human players [11].

Some other popular examples that have been used for research or industry are: AI-generative art [22], sentiment analysis tools [26], coding suggestion tools that utilize existing source code as reference such as GitHub Copilot [18] etc.

3 Methodology

To answer the outstanding research question, the method selected for research is a literature review. PRISMA workflow is the main tool that will be utilised to systematically select the papers included in the study. This flow aids researchers in bringing visualization and tracking of the screening process [19], and in the scope of this paper specifically for the sources utilised. To gather these records, a handful of websites will be used such as journal databases (Science Direct, Arxiv, Google Scholar), and search engines (Google, Bing).

When utilizing the search function in journal databases, advanced queries have been constructed to add the possibility of replicating the search process. The main terms and their synonyms were gathered and linked together with boolean logic to arrive at the first query:

> *("knowledge elicitation" OR "knowledge extraction") AND ("Large Language Models" OR "LLMs" OR "GPT" OR "Llama" OR "Gemini" OR "AI") AND ("integration" OR "applications" OR "human-computer interaction") AND ("unexplored" OR "future work" OR "new direction")*

However, this query has yielded too many entries at around 13.200 results on Google Scholar, therefore the query has been modified by removing some of the synonym keywords that were observed to output studies that weren't directly linked to the research question. Moreover, a current limitation of Science Direct also did not allow for queries with more than 8 logical operators. Therefore, the second derived query was:

> *("knowledge elicitation" OR "knowledge extraction") AND ("Large Language Models" OR "LLMs") AND ("integration" OR "applications") AND ("unexplored" OR "new direction")*

This new version has in total selected 1.640 entries in Google Scholar, 100 in Science Direct, and 16 in Arxiv. With a sum of 1.756 papers selected for review, the next step was to remove the duplicate entries and apply any remaining exclusion criteria. To find new directions that have not been explored thoroughly, and with the rapidly evolving industry around the topic of Large Language Models in mind, the exclusion criteria of removing papers older than 2023 have been applied for this research.

4 Findings

Using the query derived in the methodology section led to the identification of a large number of research papers. Following the subsequent steps of the screening process defined by the PRISMA workflow, further exclusions were made in three additional steps.

The first exclusion step involved removing studies that were not directly related to the topic. Although several studies contained the queried keywords, in some cases, the keywords were only present in the titles of references cited within the papers, which were not relevant to the context of this review. Next, some entries were excluded due to pay-walled content that could not be retrieved. Finally, the last exclusion step considered research papers that did not contribute to insights into new directions and they were therefore considered outside the scope of this study's research question. The entire process resulted in a total of 17 papers being analyzed for the literature review. The entire workflow process can be observed in figure 1.

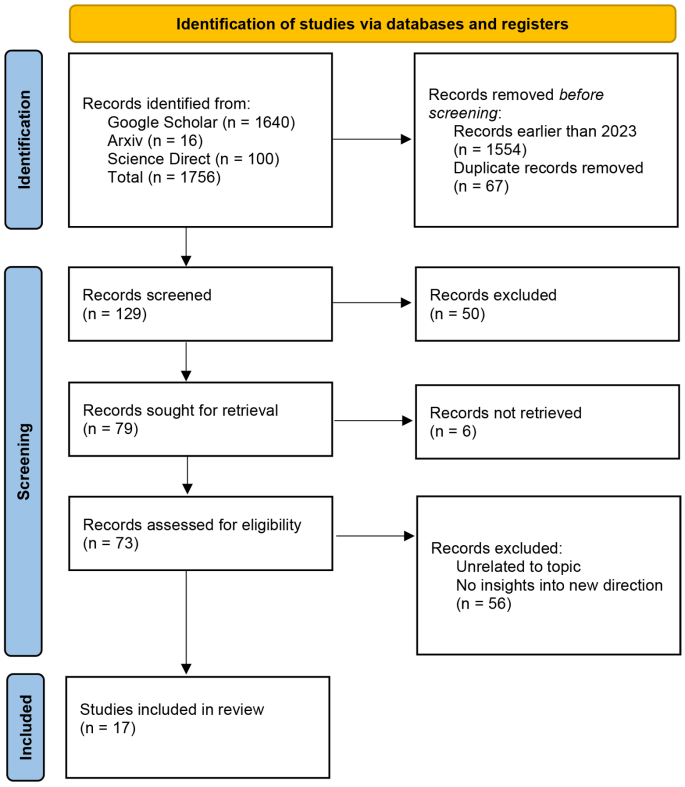

Figure 1: PRISMA workflow

After analyzing the papers included in the review, some recurring topics can be identified. To keep a clearer overview of the themes that have been used, they will be placed in distinct categories. By organizing them we can evaluate the strengths and concerns of each topic.

- Educational tools
- Knowledge curation
- Factual information
- Other topics

4.1 Educational tools

One of the most prominent topics found within the selected studies was that of educational tools. However, due to the existing link between knowledge elicitation and domain experts [25], the direction of creating tools around education is an evident direction to follow. The process of teaching is a fundamental part of education, and it usually consists of a teacher (domain expert) sharing knowledge with students (receiver of knowledge) through diverse methods. This is where a potential integration of LLMs might bring value in the domain of education. During the analysis of the multitude of research and recommended future work in this domain the most mentioned sectors that could benefit from these integrations were manufacturing [8] [30] and education in schools [2]. Moreover, other domain-specific tasks that are discussed were energy efficiency and decarbonization [33], and astronomical data [14].

The main noted benefits of integration of LLMs are their speed, user-friendliness, and logical functionality [8]. One particular concern raised specifically on this topic is the preference of students gaining knowledge from humans, rather than computer agents as the participants of user studies raise questions on reliability and safety risks [8]. This raises the question of whether the investment of resources in the industry is indeed worthwhile, and if applications can be designed around the preference for human interaction in the process of teaching. Moreover, as pointed out in the study concerning manufacturing, the research conducted has not taken place in a real-world scenario, therefore there are unknown complexities that may still be unidentified [8].

It is also notable the relevance of potential work in this area with suggestions of exploration in multi-modal LLM applications, and enhancement of a wide variety of models for domain-specific tasks [33]. However, it is important to note that in the exploratory overview of energy efficiency and decarbonization, the author only creates a correlation between the field of education and the field explored by the paper and does not perform any user studies or experiments.

Observing that most of the papers derive one specific LLM for a particular topic or domain, the call for further research into other domains in parallel could significantly contribute towards the goal of creating models for accessible and available knowledge sharing in the form of teaching and education.

4.2 Knowledge curation

Another popular topic in the research community is the intention of creating applications and models that handle the curation of knowledge for domains that hold vast amounts of data. This process involves extracting relevant knowledge from the available dataset and using it for a range of different applications such as deriving hypotheses [27]. The most prominent domains explored are the ones of bio-medicine [9], [3], [27].

In the research describing leveraging knowledge curation in generating hypotheses, the curation process consisted of prompt engineering to extract causal pairs from a total of 43.000 scientific articles from a public repository [27]. Next, after performing a ranking between probable causal concepts, the LLM was again utilized to generate a possible hypothesis between each pair of concepts [27]. This is a showcase of how knowledge curation could be leveraged for a real-world use case. However, it is important to understand the limitations of the study which mentions the difficulty of working with GPT-4 due to its lack of transparency and engineering prompts to achieve more accurate results. Moreover, the authors identify a probable 13% inaccuracy when generating the causal graph (created from the output of the knowledge curation process) which potentially affects the accuracy of the hypothesis generation [27]. This observation highlights the importance of further research into the domain of knowledge curation, as this may significantly improve the outcome of the rest of the process.

In the research on biomedical knowledge, we observe variations in the process compared to the one on psychological data. At the foundation of the paper stands the concept of knowledge distillation, which in itself, is a process of transferring information from a larger to a smaller model by processing the input data [20]. This is only achievable, however, with a reliable method of eliciting the knowledge that needs to be transferred. In the research the approach of a teacher and student LLM can be recognized with the teacher LLM (GPT-3.5) labeling data to be taught to smaller LLMs trained on the labeled data. The research identifies that this procedure leads to more accurate results with the smaller LLMs outperforming the larger ones [9]. This approach highlights a path forward for research in this domain and raises awareness of the benefits of the methodology that might apply to other domains. However, further research must ensure to address the limitation mentioned in the paper indicating the lack of experiments with GPT-4 despite initial observations being that it outperforms significantly the previous version and the lack of a more extensive dataset to be utilized for the process of distilling the knowledge [9].

Another research introduces a benchmark called LongHealth, that enables observation of how models perform on tasks regarding knowledge curation. It looks into processing patient medical records with multiple choice questions assessment for multiple LLMs. The main observation of the study showcases a lack of accuracy in the tested models regarding the information processed in a domain that requires reliability, thus concluding that the LLMs tested were not ready for real-world applications in this domain [3]. However, the main limitation of the research is the omission of a multitude of different LLMs due to a minimum requirement of 16.000 tokens and the omission of the GPT-4 model due to price consideration. Despite that, this research contributes to the objective of developing reliable LLMs that

have the potential to become a useful tool in the medical field by creating a new benchmarking tool.

Integrating LLMs for knowledge curation might lead to potential new findings and directions within the applied domain itself, as observed in the case psychology hypotheses generation [27], and potentially meaningful progress might be made with the enhancement and improvements of the models created for the applications. As LLMs are designed specifically for handling vast amounts of data, this can be one of the fields where potentially the computer can significantly improve the efficiency of research.

4.3 Factual information

Factual information is one of the challenges that current state-of-the-art LLMs struggle with. The concept of "hallucinations" is at the core of all disclaimers around applications utilizing them, due to the potential impact it might have on the end user and sensitive applications such as financial reports[7]. With this idea in mind, research seems to be actively pursued in this domain, however, meaningful improvements still need to be made. Some LLMs have been created and studied such as EntGPT [7] and TinyLlama [31] to bridge the research gap.

There are a variety of approaches when developing models around factual information as can be observed in the variety of suggested approaches in the study of EntGPT where 3 different techniques have been used (multi-step prompting, instruction tuning, and entity linking) [7]. In the aim of future work, the authors of the EntGPT study describe the intention to explore the advantages of adding entity linking on top of the entity disambiguation achieved in their paper through multi-step prompting indicating the missing knowledge in other previous research into entity linking. The research around the TinyLlama model suggests another approach for a model designed around factual information by utilizing deep learning for pattern recognition and symbolic reasoning for rule-based decision-making. This approach yielded an increase in performance in providing factual information and by design has an increase in its trust through the transparency of decision-making [31]. However, the paper acknowledges the limitation to the extent of testing the model for benchmarking without going into potential real-world use cases.

Furthermore, another study takes a different approach to contributing to research in the domain of factual information. EpiK-Eval [21] creates a benchmark for LLMs to evaluate consolidated knowledge based on the information that the LLM receives. More specifically it assesses the ability of the model to reason on the information based on a narrative it receives [21]. Benchmarks are important tools in the development of LLMs that help the research community and industry grasp the effectiveness of the model in a variety of different scenarios. Therefore, researching further on developing assessment tools on the performance and accuracy of LLMs can bring valuable insight into the future development of models and applications.

However, as indicated in the papers standing behind the previously mentioned models, there is a lot of room for improvement and future research is needed in this area to provide a better and safer experience for any end user.

4.4 Other topics

Aside from the main categories analyzed so far, a couple of other new directions have been identified scattered across literature that do not fall within a concrete category. In this subsection, we will also briefly analyze these directions as they do contain valuable insight into some interesting use cases and challenges.

Another paper performs a rapid review of potential use cases of LLMs in nursing. The key highlight of the benefits found in the rapid review is the target of relieving burden levels for medical staff during a time period when the healthcare system faces staff shortages. The outcome of the review identifies many potential directions for tools useful for nursing such as diagnostic assistance, personalized care, and multilingual support for patients as well as some suggestions of the utility in education [10]. The last two directions indicate a need for further development and improvements in applications such as chatbots. However, the topic may prove to be somewhat controversial as already some products, such as Tessa created by the National Eating Disorder Association (NEDA), have been released to the public with controversies around them leading to them being taken down due to concerns on risks and inaccurate responses potentially caused by hallucinations [12]. Just as it is highlighted in the rapid review, integrations of LLMs in the healthcare systems must be approached with a thorough understanding of the implications of legal, ethical, and privacy nature [10] as it is one of the domains with the most vulnerable category of stakeholders.

Secondly, an application that develops a dynamically evolving LLM is LLaMALoop [28] which looks into creating an LLM that has a continuous learning process with the aim of increasing response accuracy for more context-aware and user-centric experiences. This paper brings forth the idea of adapting LLM responses based on the feedback extracted from the end-user and adapting the responses accordingly. The research successfully identifies key benefits in benchmarks such as information retrieval accuracy, however, it also acknowledges trade-offs in computational performance with increased processing times[28]. The research also tackles the possibility of increasing the end-user trust in the model, however, no apparent user study was performed on this and the actual impact of this method remains generally unknown unless further research is conducted. Moreover, the need for LLMs to have a continuous learning process is also highlighted by the paper which performs a survey on the benefits of integrating code in the training data of LLMs [32]. In their paper, the key benefit of unlocking a reasoning ability for the LLM and the ability to produce more reliable intermediate steps in coding queries is identified when training LLMs with code data in their synthesis of literature [32]. However, they also suggest that a method of reinforcement learning can be an effective approach for reducing the need to provide multiple interactions toward task completion [32].

Thirdly, automatic grading tools for open-ended questions are another direction that seems to have a lot of potential. A research paper stood out exploring the potential of using LLMs for grading exams. It tests fine-tuned LLMs trained on relevant data to provide automatic grading for tests. Such a

tool aims to bring a solution to the problem of the tedious and time-consuming process of manual grading[13]. Despite the research only aiming to analyze the effects of fine-tuning a grading model, it does have significant improvements in classifying answers correctly between correct, contradictory, and incorrect for unseen questions and domains[13]. With further research such an integration could indeed achieve the target of adding efficiency to the process of grading by partially automating it, therefore further research and some experiments on a focus group of students might provide further insight into the domain.

Lastly, leveraging knowledge elicitation in LLMs for navigation is a domain that has had significant advancements in research. In the survey on advances in embodied navigation using LLMs [16] a wide variety of models developed for navigation are analyzed. Multimodality stays at the core of these LLMs as multiple types of input and training data such as text, image, and audio, need to be processed to achieve good performance in task completion. New milestones for multimodality are seemingly being achieved in products available to the public with the announcement of the GPT-4o model, however, at the time of writing this paper, the visual capability of the model is still yet to be released to the public. The capability of LLMs to actively serve as a decision-making actor is brought forth in domains such as autonomous driving [16]. However, currently bringing such systems into products is still an objective that is still not achievable due to the multiple safety and reliability concerns. Therefore, eliciting multimodal knowledge is a topic that has yet to be thoroughly researched and assessed especially concerning methods of processing the data, considering the available models and products, despite the extensive research performed so far.

4.5 Concerns

The key concerns of researchers, end-users, and industry are common ground across all applications that utilize LLMs. This is because the technology of Large Language Models themselves comes with risks that have both been publicly acknowledged such as hallucinations but also discovered in the process of development, with a variety of different attacks and even new types of attacks such as malicious prompt engineering [1]. Other concerns depicted in the literature analyzed have been in regards to the transparency of the models and applications around them [15], but also the safety and efficiency of them[8]. These concerns have been raised at a stage where the new integrations are published frequently to keep up in the potential AI race [5]. However, these implications do not mean that research and development should be completely held back, but rather that there is a need for security-centered design of applications due to the potential unidentified and unapparent risks that may lie underneath.

5 Discussion

5.1 Implications

The findings presented intend to help in bridging the gap in the knowledge available for new potential directions for knowledge elicitation integrated with LLMs. This paper provides a summary of some of the key findings of the past 1.5 years of research on integrations of LLMs for knowledge elicitation and the aim is to be a useful resource for finding research opportunities in the domain.

5.2 Research Process

An insightful topic to be addressed is the potential of the research process used in this paper. PRISMA workflow has been instrumental in keeping track systematically of the progress in the screening process of the literature. Moreover, we acknowledge that the query utilized in this research potentially can be further refined, and different queries can be utilized to explore a broader range of studies and to uncover additional relevant literature that may have been missed initially.

5.3 Future Work

While this research focused primarily on knowledge elicitation, several areas for future work could provide insight into the domain of LLMs and knowledge elicitation. Two possible directions for research stand out. Further literature reviews of gathering new directions for integrating LLMs in other domains, and further research in the directions identified as part of this paper. This study aims to encourage the industry and researchers to explore more in-depth the directions identified as part of this paper.

5.4 Limitations

The study was conducted within a tight timeframe of 10 weeks, which limited the depth and breadth of the literature review and analysis. A longer time period would have allowed for a more comprehensive exploration of the topic. The focus of the research was primarily on integrations for knowledge elicitation, however, a gap remains for other domains where LLMs can provide value. Moreover, only papers published in the English language were selected for this study. This language bias may have excluded valuable research published in other languages, potentially leading to an incomplete understanding of the worldwide research landscape.

The field of study of LLMs has rapid advancements with new models and technologies emerging frequently. Additionally, a recognizable number of research papers are in the process of peer review and are still in the process of publication. Consequently, the findings may quickly become outdated as new developments occur and additional research is published.

6 Responsible Research

It is important to acknowledge that a selection bias can affect the comprehensiveness of a literature review. Through this study, we recognize the potential for selection bias due to performing non-exhaustive research, which stems from the tools, papers, and databases utilised. To mitigate this, the process has been developed transparently, specifying the tools and the steps performed, and the criteria used for selecting the literature. This mitigation also plays a role in the effort to reduce potential reproducibility bias by ensuring that the process can be repeated with similar results. However, the exact numbers might not be able to be reproduced as the research

in the domain is continuously performed and more literature gets published.

In academic research, another important consideration has to be given to publication bias. The majority of the literature selected was observed to have positive results leaving a theoretical high probability of some directions that have yielded negative or null findings being left out.

Moreover, a limitation of this research can be correlated to a language bias. It is probable to some possible directions for research has been mentioned in literature in other languages other than English. However, to mitigate this bias the consideration for the selection mentions the limitation that only papers available in the English language have been used. This opens the possibility for future research to collect literature from more diverse sources to achieve a more comprehensive study.

7 Conclusions

To conclude, this research identifies several potential directions for both academic research and industry applications, highlighting their benefits and implications. Those directions include tools for education, knowledge curation, factual information, and other applications such as specialized chatbots, feedback loops in LLMs, automatic grading tools, and navigation. In addition to identifying the directions, this research also highlights key concerns such as transparency, safety, efficiency, and security. It also raises a recommendation for end-user and security centered designs for putting the integrations into practice.

Finally, this research provides a foundational understanding of the potential directions and benefits of integrating technologies into various domains for knowledge elicitation. Future research should continue to explore these areas, expanding the scope and incorporating a broader range of perspectives to fully realize the potential of these technologies. This common effort should be aimed at creating solutions that are not only innovative and effective, but also responsibly designed.

References

[1] Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. Securing large language models: Threats, vulnerabilities and responsible practices, 2024.

[2] Bilal Abu-Salih and Salihah Alotaibi. A systematic literature review of knowledge graph construction and application in education, 2 2024.

[3] Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL. Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. Longhealth: A question answering benchmark with long clinical documents, 2024.

[4] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using a configurable game. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1709–1719, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Stephen Cave and Seán S ÓhÉigeartaigh. An ai race for strategic advantage: rhetoric and risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 36–40, 2018.

[6] Nancy J Cooke. Varieties of knowledge elicitation techniques. *International journal of human-computer studies*, 41(6):801–849, 1994.

[7] Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. Entgpt: Linking generative large language models with knowledge bases, 2024.

[8] Samuel Kernan Freire, Chaofan Wang, Mina Foosherian, Stefan Wellsandt, Santiago Ruiz-Arenas, and Evangelos Niforatos. Knowledge sharing in manufacturing using llm-powered tools: user study and model benchmarking. *Frontiers in Artificial Intelligence*, 7, 2024.

[9] Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, and Hoifung Poon. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. 7 2023.

[10] Mollie Hobensack, Hanna von Gerich, Pankaj Vyas, Jennifer Withall, Laura Maria Peltonen, Lorraine J. Block, Shauna Davies, Ryan Chan, Liesbet Van Bulck, Hwayoung Cho, Robert Paquin, James Mitchell, Maxim Topaz, and Jiyoun Song. A rapid review on current and potential uses of large language models in nursing, 6 2024.

[11] Sihao Hu, Tiansheng Huang, Fatih Ilhan, Selim Tekin, Gaowen Liu, Ramana Kompella, and Ling Liu. A survey on large language model-based game agents. 4 2024.

[12] Nari Johnson, Sanika Moharana, Christina N. Harrington, Nazanin Andalibi, Hoda Heidari, and Motahhare Eslami. The fall of an algorithm: Characterizing the dynamics toward abandonment. 4 2024.

[13] Nazmul Kazi and Indika Kahanda. Enhancing transfer learning of llms through fine- tuning on task - related corpora for automated short-answer grading. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1687–1691, 2023.

[14] Vladyslav Kutsuruk. Astronomical data features extraction and citation prediction. 2023.

[15] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[16] Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu, Changwei Wang, Man Zhang, Li Guo, and Shibiao Xu. The development of llms for embodied navigation, 2023.

[17] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed

Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. 7 2023.

[18] Nhan Nguyen and Sarah Nadi. An empirical evaluation of github copilot's code suggestions. pages 1–5. Institute of Electrical and Electronics Engineers Inc., 2022.

[19] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *bmj*, 372, 2021.

[20] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Chapter 8 - knowledge distillation. In Alexandros Iosifidis and Anastasios Tefas, editors, *Deep Learning for Robot Perception and Cognition*, pages 165–186. Academic Press, 2022.

[21] Gabriele Prato, Jerry Huang, Prasannna Parthasarathi, Shagun Sodhani, and Sarath Chandar. Epik-eval: Evaluation for language models as epistemic models, 2024.

[22] Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. Diffusiongpt: Llm-driven text-to-image generation system. 1 2024.

[23] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, 1 2023.

[24] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. 8 2017.

[25] Nigel R Shadbolt, Paul R Smart, J Wilson, and S Sharples. Knowledge elicitation. *Evaluation of human work*, pages 163–200, 2015.

[26] Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. Sentiment analysis through llm negotiations. 11 2023.

[27] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating psychological hypothesis generation with ai: Large language models meet causal graph. November 2023.

[28] Hsiao-Ching Tsai, Chih-Wei Kuo, and Yueh-Fen Huang. Llamaloop: Enhancing information retrieval in llama with semantic relevance feedback loop. 2023.

[29] L. von Ahn. Games with a purpose. *Computer*, 39(6):92–94, 2006.

[30] Liqiao Xia, Chengxi Li, Canbin Zhang, Shimin Liu, and Pai Zheng. Leveraging error-assisted fine-tuning large language models for manufacturing excellence. *Robotics and Computer-Integrated Manufacturing*, 88:102728, 2024.

[31] Xingyu Xiong and Mingliang Zheng. Integrating deep learning with symbolic reasoning in tinyllama for accurate information retrieval, 01 2024.

[32] Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R. Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, Heng Ji, and Chengxiang Zhai. If llm is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents, 2024.

[33] Liang Zhang and Zhelun Chen. Opportunities and challenges of applying large language models in building energy efficiency and decarbonization studies: An exploratory overview, 2023.