Leveraging large language models in games with a purpose for enhanced knowledge elicitation

Tommy Hu[1]
Supervisor(s): Ujwal Gadiraju[1]**, Shreyan Biswas**[1]
[1]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: Tommy Hu
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Shreyan Biswas, Ricardo Marroquim

Abstract

The process of knowledge elicitation is crucial to the field of artificial intelligence because of the lack of data on commonsense knowledge. This paper explores the potential of using large language models (LLM) to enhance knowledge elicitation in games with a purpose (GWAP). By analyzing the capability of LLMs to play games and generate game content, this research shows how LLMs can increase accessibility, efficiency and engagement in GWAPs for knowledge elicitation. The findings show that LLMs can play games with great performance and show much promise in generating game content that facilitates knowledge elicitation. Through a comprehensive literature survey, this research highlights potential ways of enhancing knowledge elicitation in GWAPs using LLMs and offers recommendations for future research in this field.

1 Introduction

Most humans possess commonsense knowledge, but according to the Gricean maxims this knowledge is typically omitted in written or oral communication [6]. Data regarding commonsense knowledge is however very important in the field of artificial intelligence since it is needed to truly understand human behavior [4]. To obtain this data we need to extract this commonsense knowledge from humans and turn it into a machine-readable format. One method for eliciting this knowledge is the utilization of games with a purpose (GWAP), which aims to provide humans with an enjoyable experience in the form of a game while collecting useful data.

Since the release of the first GWAP for collecting commonsense knowledge, Verbosity [1], many different GWAPs have been created such as FindItOut which is a multiplayer GWAP between two human players that aims to collect positive and negative, generative and discriminative knowledge [2]. Many of these GWAPs have been very successful in eliciting commonsense knowledge, however, there is still room for improvement in diversifying the types of knowledge elicited and making the games more accessible, which could increase the amount of knowledge elicited [2]. A potential way to address these improvements is through the usage of large language models (LLM) that act as intelligent agents within the game environment by understanding and generating text. This paper researches the possibilities of utilizing LLMs to enhance the process of knowledge elicitation using GWAPs by answering the following questions:

- What type of games can be utilized for knowledge elicitation
- How have LLMs been used to play games and how well did they perform
- How can we utilize LLMs to generate game content that facilitates knowledge elicitation

By answering these questions we present the following contributions:

- Demonstrating the potential of LLMs to enhance knowledge elicitation in GWAPs by replacing human players and generating game content.
- Highlighting what type of games can be used for knowledge elicitation and what characteristics of these games allow for knowledge elicitation.
- Providing insights into the design of future GWAPs, leveraging LLMs to create more engaging and targeted knowledge elicitation experiences.

The paper is structured in the following way: First, the related work discusses the previous work on this topic in detail after which the research methodology explains the methods for finding the papers included in the survey. Following the methodology the background explains relevant background concepts including the relevance of the research. Subsequently, the findings are presented and discussed followed by a section on the limitations of the research. Lastly, the conclusion summarizes the research and answers the research questions.

2 Related Work

Much work has already been done in the field of knowledge elicitation using games with a purpose (GWAP). The first GWAP that did this is Verbosity which collected descriptions of certain concepts [1]. Many other GWAPs have been created that collect different types of knowledge such as Common Consensus [10], which collected knowledge regarding human goals, and the 20 Questions game [13], which collected positive and negative assertions. One of the

more recent GWAPs for knowledge elicitation is FindItOut, which collected many different types of knowledge and also characterized the types and quality of the knowledge [2]. With all this previous work we have seen many different ways of eliciting knowledge using GWAPs and with this information, we aim to look into methods to enhance this process by utilizing LLMs.

3 Methodology

To answer the research questions a literature survey was conducted using papers found on Google Scholar, Scopus and in the references of other papers. The set of papers for this survey consists of papers regarding games with a purpose (GWAP) for knowledge elicitation and large language models (LLM) being used to play games or generate game content. Tools like Scopus were used to find a lot of these papers by searching the words "large language model" and "game" within the title, abstract and keywords of papers, which resulted in an initial list of 599 papers. To further refine the results other specific filters were used like restricting the subject area to Computer Science, document type to conference papers and only selecting papers with the keyword "Language Model". From the resulting list of 120 papers regarding the usage of LLMs in games, 4 papers were selected by assessing the relevance of the title and abstract. To obtain papers regarding the process of knowledge elicitation in games, 3 papers were selected from the references of a very recent game for knowledge elicitation, FindItOut [2].

4 Background

4.1 Knowledge Elicitation

Knowledge elicitation is a process in knowledge engineering that aims to extract knowledge from humans and transform it into a machine-readable format. This knowledge is crucial for the development of artificial intelligence systems since it is believed that this knowledge is needed to create truly intelligent systems [1]. Knowledge can vary from explicit to tacit or from general to specific, however, many previous works lack an in-depth characterization of the collected knowledge [9]. Knowledge elicitation methods are needed because there is a lack of data regarding knowledge. This lack of data exists because most humans already have this commonsense knowledge which removes the need to include it in written communication [6].

4.2 Games with a Purpose

One method of eliciting knowledge is through the use of games with a purpose (GWAP). This method involves one or more humans using their commonsense knowledge to play a game and by recording and analyzing the actions of the humans knowledge can be extracted from the gameplay. One example of this is the game FindItOut where two players are assigned a card with an image on it and by asking and answering questions they need to find out what the image is on the opposing player's card [9]. Using GWAPs is a great way of eliciting knowledge since providing a fun game experience can increase the motivation of a participant to contribute their knowledge which results in a higher participation rate and more knowledge elicited.

4.3 Large Language Models

One possible way to enhance the process of knowledge elicitation is through the usage of large language models (LLM). Large language models are based on neural networks that utilize deep learning techniques to understand and generate natural language. Some LLMs have been shown to possess consistent world knowledge on various categories allowing them to achieve strong performance on natural language processing tasks [3]. Despite their high performance on natural language tasks, LLMs could also be used in non-linguistic tasks due to their high versatility [14]. These capabilities of LLMs could allow for them to be used to enhance the process of knowledge elicitation in GWAPs by acting as players in the game or generating game content.

5 Findings

5.1 Knowledge elicitation in GWAPs

Many games with a purpose (GWAP) have been created with the goal of collecting large amounts of knowledge. Verbosity was the first GWAP that implemented this idea in the form of a two-player collaborative game [1]. In this game, one of the players is assigned the role of "Narrator" while the other player

receives the role of "Guesser". The Narrator gets a secret word which the Guesser must try to guess based on hints given by the Narrator. The Narrator can give hints by choosing a sentence template that contains a blank space to be filled in and filling in the blank such that the sentence describes the secret word. An example of this could be the sentence template: "It contains a ..." which could be filled in with the word "keyboard" to describe the secret word "laptop". After the Guesser tries to guess the word based on the hint the Narrator can tell the Guesser whether the guess is "hot" or "cold" to guide the Guesser. By analyzing the sentences created by the Narrator Verbosity collects commonsense knowledge about a large quantity of words.

Common Consensus is another multiplayer GWAP that tries to collect knowledge regarding goals that motivate human behavior [10]. The gameplay of Common Consensus revolves around the players providing as many possible answers to a trivial, open-ended question. The game uses six question templates where a goal could be filled such that knowledge can be elicited from the answers to the question. An example of this would be the question: "What are some things you would use to watch a movie?" where possible answers would be "dvd", "television" or "tv".

Unlike the previous 2 GWAPs, the 20 Questions game is a singleplayer GWAP focused on eliciting the most salient characteristics of particular concepts [13]. This game asks the player 20 questions regarding a particular concept which the player has to answer with yes or no. This gameplay system allows for the collection of negative assertions that are underrepresented in commonsense knowledge bases [13]. Behind the scenes, the 20 Questions game generates questions based on already present knowledge which intend to identify a small cluster of objects that can be compared to the object in question.

Lastly, FindItOut is a competitive two-player game that can collect many different types of knowledge [2]. FindItOut presents both players with multiple cards containing semantic concepts. The goal of the game is to guess the other player's card by formulating questions using sentence templates. Players can answer "yes", "no", "maybe" or "unclear" when responding to questions which also allows for the elicitation of negative knowledge.

All these games have various gameplay mechanics allowing for the elicitation of different types of knowledge. GWAPs for knowledge elicitation can vary from singleplayer to multiplayer or from collaborative to competitive. One characteristic that is often used for knowledge elicitation is the usage of sentence templates which the player must fill in using their commonsense knowledge, this can be seen in Verbosity and FindItOut. The 20 Questions game and FindItOut both have the player answer questions with "yes" or "no" to elicit positive and negative assertions. Lastly, Common Consensus utilizes open-ended questions to collect a wide range of answers. Yet, one thing these games have in common is the generation of sentences that form the basis of the elicited knowledge. In some of the games sentences are generated by the player while in other games they are generated by the game, these sentences can take the form of questions or even descriptions of a certain topic. The generation of sentences followed by the response of the player forms the core of the knowledge elicitation process using GWAPs.

5.2 LLMs playing games

Large language models (LLM) are capable of understanding and generating text which allows them to act as agents in text-based games. Since GWAPs for knowledge elicitation have text-based gameplay LLMs could also be used to play these games. One game that has been played by LLMs is the 20 Questions game which requires a significant amount of world knowledge [3]. In this game the LLM needs to answer various questions regarding a specific entity such as a keyboard. To answer these questions the LLM needs to possess knowledge regarding the shape, composition or purpose of a keyboard. This task would be easy for humans but could be difficult for LLMs. The results of LLMs playing the 20 Questions game showed that smaller language models lack the world knowledge to perform well in this game, however, other models such as GPT-3 performed very well on many categories [3]. In the experiments, LLMs were tested in a zero-shot setting where they only received a textual description of the task. In this setting, GPT-3 achieved an accuracy of 81.3% which was increased to 87.9% by providing four examples in addition to the zero-shot setup [3]. These results show that state of the art LLMs are capable of playing games that require a large amount of world knowledge with high accuracies.

Another implementation of LLMs playing games is the clembench framework, which allows LLMs to play various Dialogue Games while evaluating the performance of the models [9]. A Dialogue Game is de-

scribed as a constructed activity with a clear beginning and end, in which players attempt to reach a predetermined goal state primarily by means of producing and understanding linguistic material [12]. To implement Dialogue Games with LLMs as players clembench uses a programmatic Game Master who keeps track of the game state and ensures players are correctly following the rules by parsing their inputs. One of the Dialogue Games that LLMs played in this framework is the Taboo game where a player must describe a concept to another player without mentioning the concept name. In the experiments, LLMs played the Taboo game against themselves and were evaluated on the number of games they played to completion and the quality of the gameplay. The results once again showed that the GPT models achieved the best performance with GPT-4 having played 95% of games to completion with a quality score of 76 [9]. These results indicate that LLMs are capable of playing high quality games of Taboo and are likely capable of replicating this performance in similar GWAPs such as Verbosity.

Lastly, a promising use for LLMs is the role of generative agents in an immersive role-playing environment. Research has been done into generative agents that simulate human behavior such as waking up, cooking breakfast and heading to work [11]. In this research, multiple generative agents were instantiated to populate an interactive sandbox game world, Smallville, and simulate day to day life. These agents produce believable individual and emergent social behaviors by utilizing LLMs in combination with computational interactive agents [11]. The identity and relations of each agent are depicted by one paragraph of natural language description and the agents output their current action in a natural language statement at each time step of the sandbox engine [11]. By having human players interact with the sandbox environment and generative agents via natural language these types of games could potentially allow for knowledge to be elicited. Since humans often use commonsense knowledge to make sense of everyday situations [7], this knowledge could be elicited by having the game present various everyday situations to human players and analyzing the actions of the human. Interactions with other agents or humans could also allow for knowledge regarding social relationships to be elicited.

LLMs have demonstrated great performances playing games like the 20 Questions game and the Taboo game. Both of these games are able to be utilized as GWAPs for knowledge elicitation making it possible for LLMs to be utilized in those GWAPs. LLMs could also be used for knowledge elicitation in a more complex game environment such as Smallville. Role-playing games like Smallvile have been shown to be extremely popular among the gaming community, one example of this being the Sims franchise which sold over 200 million copies [5]. Incorporating knowledge elicitation in a role-playing game could allow for a much larger amount of players playing the game resulting in a larger quantity of knowledge.

5.3 LLMs generating game content

In the previously discussed game, Smallville, the generative agents perform actions based on the context that was provided to them including their identity and relationships. The process of creating the identities and relationships of these agents is something that could be enhanced through the usage of LLMs. To utilize games like Smallville for knowledge elicitation the game needs to create situations that facilitate knowledge elicitation, which could also be supported by LLMs. Utilizing LLMs to generate game content could be a great way to enhance the development process of games.

One way LLMs have been used for this is through an end-to-end framework, SceneCraft, that generates a story with corresponding dialogues based on various details given by the author [8]. This framework generates the story, chooses relevant emotes and gestures to be used by the characters and generates all the code to be directly used by the game engine. The generated story contains branched narratives allowing the user to influence the story through their actions. The framework was evaluated by 9 participants with varying game development experiences. The evaluations showed that users had favorable opinions regarding the ease of use, satisfaction and adaptability of SceneCraft, however, there is still room for improvement on the creativity and character engagement side of things [8]. These results show that LLMs are capable of automating the generation of character interactions based on instructions from the author. By writing instructions to facilitate knowledge elicitation the generated content could be used to enhance the process of knowledge elicitation in games.

Aside from generating natural language, LLMs have been utilized in the realm of procedural content generation in games. Research was done into using the GPT-2 and GPT-3 models to generate game

levels for the puzzle game Sokoban [14]. This research showed that even though these models were trained to solve linguistic tasks, they are also capable of more specialized tasks like generating valid game levels. Using LLMs in this way allows levels to be generated with specific characteristics to adapt them to the player's knowledge and skills. In the experiments, the levels were generated in a few-shot setting where they were given example levels with annotations like the length of the solution. The results showed that different models of GPT-2 were all able to reliably generate novel, playable and diverse levels without copying them from the training data. Since the levels of puzzle games must have a solution it would be difficult to elicit factual knowledge using these levels, however other types of knowledge could be elicited. By analyzing the way players solve these levels knowledge could be elicited regarding the decision-making process and problem-solving approaches of humans. Incorporating knowledge elicitation in puzzle games could be a way to make the experience more enjoyable for players while collecting knowledge.

These results show that LLMs are highly versatile and are capable of generating game content based on specific characteristics. This strength of LLMs could prove useful in enhancing the knowledge elicitation process in current GWAPs and in new directions of GWAPs. The controllability of the generated content could be used to elicit knowledge with specific characteristics and also adapt the content to the capabilities and knowledge of the player. LLMs could also be used to generate content in more complex games, which could provide players with more satisfaction allowing for a larger player base.

6 Discussion

6.1 Implications

The findings from this research have significant implications in the field of knowledge elicitation using games with a purpose (GWAP). This research explores the possibility of utilizing large language models (LLM) to play games. This could increase the accessibility of many previous GWAPs that required multiple humans for a single game by replacing human players with LLMs. This increase in accessibility allows for more humans to play the game resulting in a larger quantity of knowledge elicited. The generation of game content utilizing LLMs could also speed up or even automate parts of the development process of GWAPs while collecting targeted knowledge. These methods show much promise in enhancing the process of knowledge elicitation using GWAPs

6.2 Future Work

Some open issues and new questions have been identified through this research. One question that arises from LLMs playing GWAPs for knowledge elicitation is what to do with the data generated by the LLMs. This data could be chosen to be ignored, however, research could be done into possible use cases for this data.

Another area for future research is the utilization of more complex games like role-playing games for eliciting knowledge. These games have the potential to attract a large amount of players, however, incorporating knowledge elicitation into these games is not as straightforward and could be researched further.

7 Limitations

Reflecting on the ethical aspects of this research, it is important to address certain limitations. The selection of papers from databases such as Google Scholar and Scopus may lead to a selection bias since these databases do not contain all papers. The use of specific keywords and filters could also exclude results that contain valuable information. Lastly, the rapid advancements in the development of large language models could mean that some of the sources are already outdated.

It is also important to address the reproducibility of the methods used in this research. Despite documenting the keywords and filters that were used to find the sources, the final selection of papers was selected through a subjective assessment of the relevance of the title and abstract. This subjective assessment could affect reproducibility since different researchers could have different opinions regarding the relevance of a paper.

8 Conclusions

This paper researches how large language models (LLM) could be used to enhance knowledge elicitation in games with a purpose (GWAP). We analyzed the type of games that could be used for knowledge elicitation and what some of the key characteristics

are that allow for knowledge to be elicited. Using this information, we researched how LLMs have been used to play games and showed that they are capable of playing GWAPs for knowledge elicitation. We also researched how LLMs have been used to generate game content and showed that they can be used to generate content that facilitates knowledge elicitation in GWAPs. With these findings, we show ways LLMs can be used to enhance the process of knowledge elicitation in GWAPs

References

[1] Luis Von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: A game for collecting common-sense facts, 2006.

[2] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using a configurable game. pages 1709–1719. Association for Computing Machinery, Inc, 4 2022.

[3] Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. Is it smaller than a tennis ball? language models play the game of twenty questions, 2022.

[4] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence, 9 2015.

[5] Elise Favis. How the sims navigated 20 years of change to become one of the most successful franchises ever. *The Washington Post*, 2020.

[6] H. P. Grice. Logic and conversation. pages 41–58. 1975.

[7] Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro Szekely. Dimensions of commonsense knowledge. 1 2021.

[8] Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. Scenecraft: Automating interactive narrative scene generation in digital games with large language models, 2022.

[9] Tian Liang, Zhiwei He, Jen tse Huang, Wenxuan Wang, Wenxiang Jiao, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Leveraging word guessing games to assess the intelligence of large language models. 10 2023.

[10] Henry Lieberman, Dustin A Smith, and Alea Teeters. Common consensus: a web-based game for collecting commonsense goals, 2007.

[11] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. Association for Computing Machinery, Inc, 10 2023.

[12] David Schlangen. Dialogue games for benchmarking language understanding: Motivation, taxonomy, strategy. 4 2023.

[13] Robert Speer, Jayant Krishnamurthy, Catherine Havasi, Dustin Smith, Henry Lieberman, and Kenneth Arnold. An interface for targeted collection of common sense knowledge using a mixture model. pages 137–146, 2009.

[14] Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. Level generation through large language models. Association for Computing Machinery, 4 2023.