Types of Knowledge Elicited from Games With A Purpose Using Large Language Models

Exploring Collaboration between AI Techniques and Human-Centric Game Designs

Wout Burgers[1]

Supervisors: 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: Wout Burgers (5329868)
Final project course: CSE3000 Research Project
Thesis committee: Ujwal Gadiraju, Shreyan Biswas, Ricardo Marroquim

Abstract

This research investigates the types of knowledge that can be elicited through the integration of Large Language Models (LLMs) into Games With A Purpose (GWAPs). By using a literature survey using the PRISMA framework, we synthesize findings from different studies to find patterns and gaps in the existing research. The survey focuses on the utilization of LLMs within GWAPs and examines how these models diversify the types of knowledge elicited. The findings indicate that LLMs significantly enhance the knowledge elicitation capabilities of GWAPs, transforming them into more interactive and effective tools. We found that different types of knowledge can be elicited, such as contextual insights, factual information, semantic associations and experimental knowledge.

1 Introduction

Large Language Models (LLMs), the most visible of the many recent advances in artificial intelligence, have added technological solutions on a different level [7]. As an example, LLMs in OpenAI's GPT series have been successfully applied to various works, from the production of text to enhancing the translation service in several languages [5]. Meanwhile, Games with a Purpose (GWAPs) have become an especially powerful approach within the whole set of methods used to understand human cognitive abilities so that problems currently beyond the abilities of a normal computer can be solved. This kind of concept was first initiated in the idea of von Ahn and Dabbish in their work on reCAPTCHA [22]. Core to the success of GWAPs is the ability to convert tedious kinds of annotation activities into, instead, entertaining forms that attract extensive kinds of participation.

Generally, both LLMs and GWAPs have proven individually in some applications, but very little effort has been made to put both technologies together, especially in how LLMs should be adapted to maximize the extraction of knowledge within a gamified environment. This gap is attempted to be plugged by performing a systematic literature survey of the kinds of knowledge that can be extracted directly through the inclusion of an LLM in GWAPs.

Understanding the interaction between LLMs and their human users within GWAPs and using that knowledge to optimize this interaction is important. This importance is there because there is a pressure that large, highly accurate annotated data is needed to train better machine learning models. Earlier work has shown that LLMs can generate human-like responses as part of their training data, which suggests that such models may be able to largely replace the need for human expertise in training other models [8]. Additionally, the interactivity within GWAPs is useful for the research of the collaboration between humans with AI. This is something that has been identified as a field that has to be developed seriously to further the usability and acceptability of AI within our society [13].

By learning from existing LLM implementations wihtin GWAPs that have been used in contexts other than data collection, such as the ”Word Ladders” [1] mobile application . This application is embedding knowledge collection together with engagement, drawing from semantic data, and aligning it with user engagement through gamification [3]. This paper explores similar strategies that can be used in GWAPs to arrive at eliciting a broader set of knowledge types. The final goal here is to understand not only what knowledge can be extracted but also how it can be leveraged in a way that improves the usability of AI technologies and the data collection process.

To address this gap, our research focuses on the main research question: **What type of knowledge can be elicited using LLMs in GWAPs?** This question aims to look at the potentials of LLMs to enhance the effectiveness of GWAPs in collecting diverse types of knowledge. By systematically reviewing existing literature, we try to identify how LLMs can be adapted to improve user engagement and the quality of data collected in gamified environments. We hypothise that the general power of LLMs will basically augment the data collection power of GWAPs both in terms of user engagement and collected data quality [24].

To address the research question, we have formulated several sub-questions that help to delve deeper into specific aspects of this research question. These sub-questions aim to looks at various perspectives of the interaction between LLMs and users within a gamified environment:

1. What kinds of semantic associations do players reveal when interacting with LLMs in GWAPs?
2. What are the challenges and limitations of using LLMs to elicit specific types of knowledge in a GWAP setting?
3. How does the player's demographics (age, educational background, etc.) influence the quality and type of knowledge elicited by LLMs in GWAPs?

By addressing these sub-questions, our research aims to provide a detailed understanding of the multifaceted interactions between LLMs and users in GWAPs. This will elucidate the types of knowledge that can be elicited.

The following sections will discuss LLM and GWAP literature and how they individually contribute to AI and data collection. Subsequently, we will detail the methodology of our research. Then we explain our findings, and potential importance. We will then progress into the discussion with the limitations and future work.

2 Methodology

In this chapter, we show the methodology used to address our research question and its sub-questions. The methodology is divided into three main sections: an explanation of the PRISMA workflow, the motivation for using this method, and the process of arriving at the final set of papers included in our literature survey.

[1] Link to application: https://play.google.com/store/apps/details?id=it.synesthesia.abstraction&hl=en_US

2.1 PRISMA Workflow

To ensure a comprehensive and transparent literature survey, we adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. PRISMA is a well-established methodology designed to improve the reporting and systematic process of reviewing literature, making it particularly suitable for synthesizing research findings across multiple studies. The PRISMA workflow involved the four key steps: identification, screening, eligibility, and inclusion [15].

2.2 Motivation for Using PRISMA

The choice of the PRISMA framework was motivated by its structured approach and widespread acceptance in systematic reviews. PRISMA enhances the transparency and reproducibility of the survey process, making it easier to follow and replicate [15]. This method is particularly beneficial for our study as it provides a clear, step-by-step approach to identifying, screening, and selecting relevant studies. Which is ensuring that our literature survey is both comprehensive and unbiased.

We conducted a literature survey to systematically explore the existing body of work on the integration of LLMs in GWAPs. The PRISMA framework allowed us to organize our search strategy, apply good selection criteria, and ensure that a lot of relevant studies were considered. This structured approach is crucial for synthesizing findings from diverse studies and drawing meaningful conclusions about the research question, namely the types of knowledge that can be elicited.

2.3 Arriving at the Final Set of Papers

The process of arriving at the final set of papers involved multiple stages of selection to ensure that only the most relevant and high-quality studies were included. Relevant studies are those that directly address the integration of LLMs into GWAPs and the types of knowledge elicited through this integration. High-quality studies are characterized by their methodology, data, and clear, well-supported findings. By focusing on studies that meet these criteria, the research ensures that the conclusions drawn are based on the most credible and pertinent evidence there is. We used the following query to find the academic papers:

(Large Language Models OR LLM) AND (Games With A Purpose OR GWAP*) AND (gamification OR interactive) AND knowledge elicitation AND types of knowledge AND (user engagement OR data quality OR semantic associations OR demographics) AND (knowledge collection OR data collection) AND (user interaction OR user studies) NOT (entertainment AND gaming industry AND marketing AND advertising)*

The query resulted in a set of around 7890 papers. The final pool of around 30 papers relevant to our research was arrived at after using the PRISMA workflow. We categorized the studies based on their focus areas, methodologies and findings allowing us to analyze the data and extract relevant insights. The existing research was covered by this process, providing an understanding of how LLMs can enhance knowledge elicitation in GWAPs.

By following this process, a thorough and reliable literature survey was ensured, laying a robust foundation for addressing the primary research question and guiding future research in this innovative field.

3 Background

This section provides an overview of the technologies of the research, which are the LLMs and GWAPs. We discuss the evolution and capabilities of LLMs, explaining their role in processing and generating natural language on a large scale. Moreover, we look at the concept of GWAPs which leverage human intelligence tasks within a game-like application to solve problems that are difficult for machines. Lastly, we discuss how the LLMs are integrated within the GWAPs.

3.1 Large Language Models (LLMs)

Large Language Models, such as the OpenAI ChatGPT series, have significantly advanced the capabilities of natural language processing (NLP) systems. This means that these models, which are equipped with tens of billions of parameters and holding a neural network architecture of deep layers, are able to produce accurate and even contextually relevant text based on the given prompt [16]. These models have an architecture that has been inspired by the Transformer model and making training over large datasets easy [20]. What is certain, is that LLMs have been successful in all of the various tasks it has had to execute, such as text completion, translation, or content generation. This is a very powerful tool also suitable for academic research and practical applications in anything from tech to customer service industries [8].

3.2 Games With A Purpose (GWAPs)

Humans can be directed to applications to derive solutions within GWAPs, an approach that uses the power of computation to solve challenging problems for the world. This was popularized by Luis von Ahn, with projects like reCAPTCHA, which used humans in order to digitize books while providing CAPTCHA services [21]. GWAPs are designed to be funny and maintaining constant participation. This is needed in tasks requiring human effort, like image annotation, language translation, and data validation. GWAPs are so-called dual-natured; they are both entertaining and utilitarian, which makes them great media to include the public in tasks important for scientific research and data collection [23].

3.3 Integration of LLMs in GWAPs

The integration of LLMs within GWAPs is used to enable the human cognitive skills in the interaction with machine-learning capabilities. The capability of LLMs to produce human-like text could be used for a better design of the GWAP, making it more attractive and useful to apply human intelligence to help close the gap between machine capabilities and human-level capabilities.

For example, the future generation of game content can use LLMs to befit user skill level and interest. Therefore, it can help to improve user involvement and quality of collected data. Moreover, the use of LLMs through GWAPs allows participants to explore much more complex semantic relationships. These features find particular strength in pedagogical applications. Where the learning process is enriched both in quality, with thanks to the interactive and adaptive feedback mechanisms developed with LLMs [17]. The applications are not limited to pedagogical functions, as they are important for data collection but also increase the scope of influence that the GWAPs have on older data annotation tasks [3].

4 Related Work

The integration of Large Language Models (LLMs) into Games With A Purpose (GWAPs) for knowledge elicitation builds on research in natural language processing, human-computer interaction, and gamification. This section reviews the most relevant existing work, highlights their contributions, and identifies unanswered questions.

Bolognesi et al. introduced ”Word Ladders,” a mobile application designed as a GWAP to collect semantic data. This study explores how LLMs can facilitate the game by eliciting specific types of word associations from players, contributing to data collection efforts in semantic research [3]. However, the paper focuses on semantic data and does not fully explore other types of knowledge that could be elicited using LLMs in GWAPs.

Shin et al. presented AUTOPROMPT a method using automatically generated prompts to elicit factual knowledge from masked language models. This technique demonstrated the efficiency of using strategic prompts to enhance knowledge extraction from LLMs [19]. While AUTOPROMPT showcases the potential of LLMs in knowledge elicitation, it does not specifically address the gamified environments of GWAPs, leaving an open question on how these methods could be adapted for interactive games.

Luis von Ahn and colleagues developed reCAPTCHA, an important and a much used GWAP that utilized human effort to digitize books while simultaneously providing CAPTCHA services [23]. This work laid the foundation for using gamification in practical tasks, demonstrating the dual benefit of entertainment and utility. Despite its success, reCAPTCHA primarily focused on text digitization, and its methods for eliciting other forms of knowledge through LLMs remain unexplored.

Balayn et al. presented a configurable game designed to elicit diverse knowledge from players, highlighting the potential of GWAPs in gathering a wide range of data [1]. This study provides a comprehensive framework for knowledge elicitation but does not delve deeply into how LLMs can be integrated into such games to enhance the quality and diversity of collected knowledge.

Researchers have achieved significant milestones in both LLMs and GWAPs individually. For instance, Radford et al.'s work on improving language understanding by generative pre-training has significantly advanced NLP capabilities, laying the groundwork for subsequent LLM applications in various domains [16]. On the GWAP side, projects like reCAPTCHA have proven effective in leveraging human cognitive abilities for tasks that are challenging for machines, such as image labeling and data annotation [23]. These efforts have demonstrated the potential of gamification in engaging large numbers of participants to contribute to meaningful tasks.

Despite these advances, several questions remain unanswered. How can LLMs be optimally integrated into GWAPs to enhance the quality and breadth of knowledge elicited? What are the specific challenges and limitations of using LLMs in a gamified setting, particularly in terms of user engagement and data quality? Furthermore, how do player demographics, such as age and educational background, influence the effectiveness of LLMs in GWAPs?

By addressing these questions, our research aims to fill the existing gaps and provide a detailed understanding of how LLMs can be leveraged in GWAPs to elicit diverse types of knowledge. This will not only advance the field of AI but also enhance the efficacy of gamified approaches to data collection and user engagement.

5 Eliciting Knowledge from LLM-Enhanced GWAPs

This chapter presents the detailed findings of our research, addressing the main research question and sub-questions through a systematic literature survey. Our analysis draws on insights from various studies to understand the integration of Large Language Models (LLMs) into Games With A Purpose (GWAPs) for knowledge elicitation.

5.1 Findings

The integration of LLMs into GWAPs facilitates the elicitation of a broad spectrum of knowledge types. The primary types of knowledge elicited include semantic associations, factual information, commonsense knowledge, and contextual insights. These findings are supported by studies such as ”Word Ladders,” which successfully used LLMs to collect semantic data through gameplay, demonstrating the potential of LLMs to enhance the quality and depth of data collected in gamified environments [3]. To create a better idea on the research, all the findings to each of the sub-questions will be summarized. Then the findings will be synthesized to create our final conclusions.

Sub-Question 1: How do interactions of players with LLMs in GWAPs reveal semantic associations?

Semantic associations are a crucial type of knowledge that can be effectively elicited through the integration of LLMs in GWAPs. These associations help in understanding how users perceive and relate different concepts, which is vital for various applications, including language processing and AI training.

One example is the ”Word Ladders” game, where players interact with an LLM to generate semantic relationships between words. The LLM facilitates user engagement by suggesting prompts and responding to player inputs, and collecting valuable semantic data. This interaction helps uncover semantic associations that traditional methods might miss [3].

The study of semantic associations through LLMs in GWAPs reveals patterns in user behavior, such as relationships between words that are contextually specific. This ability to show nuanced semantic connections is enhanced by the LLM's capacity to understand and generate human-like text [16]. The adaptive nature of LLMs allows them to refine prompts based on player responses, leading to richer data collection.

Additional research supports the effectiveness of LLMs in eliciting semantic associations. For example, He et al. explored how LLMs can mimic human-like translation strategies, revealing that LLMs can capture semantic nuances in different languages [11]. Also, Shin et al. introduced AUTOPROMPT which uses automatically generated prompts to elicit factual knowledge, showing the efficiency of the strategic prompts [19].

Moreover, the ability of LLMs to engage users in participating in GWAPs and collect high-quality semantic data is not limited to lexical associations. Studies such as Riedl and Zook's work on AI for game production highlight the potential for LLMs to enhance data collection in interactive environments [17]. Mikolov et al. also emphasize the importance of word representations in vector space for understanding semantic associations, which is fundamental to the usage of LLMs [14].

Sub-Question 2: What are the challenges and limitations of using LLMs to elicit specific types of knowledge in a GWAP setting?

Several challenges and limitations were identified in the use of LLMs for knowledge elicitation in GWAPs.

While LLMs can generate human-like responses, the quality of the data elicited can vary. For instance, studies have indicated that LLM-generated data may sometimes include noise or irrelevant information, which can affect the accuracy of the collected knowledge [19]. This issue is shown by the tendency of LLMs to produce good-sounding information but factually incorrect information, known as ”hallucinations” [2].

Maintaining high levels of user engagement is critical for the success of GWAPs. LLMs can enhance engagement by making the game more interactive and responsive. However, if the LLMs´ responses are not engaging enough users may lose interest which is impacting the overall effectiveness of the knowledge elicitation process [1]. Moreover the uniqueness of interacting with an LLM might wear off quickly leading to less user motivation over time [21].

The integration of LLMs into GWAPs requires computational resources and technical expertise. This can be a barrier for smaller research teams or projects with limited funding. Additionally, the complexity of training and fine-tuning LLMs for specific tasks can pose technical challenges [17]. Even with access to computational resources the process of ensuring that the LLMs are aligned with the specific objectives of the GWAP can be time-consuming and requires specialized knowledge [4].

Furthermore, ethical considerations such as bias in LLMs can affect the inclusivity and fairness of GWAPs. Furthermore, ethical considerations such as bias in LLMs can affect the inclusivity and fairness of GWAPs. For instance, a study by Sheng et al. demonstrated that LLMs could generate biased language, such as associating certain professions with specific genders. Such biases in LLM responses can influence the inclusivity of GWAPs by perpetuating stereotypes and excluding or misrepresenting certain demographic groups [18]. Ensuring that models are trained on diverse and representative data is crucial to mitigate these biases and enhance the fairness of GWAPs [9].

Sub-Question 3: How does the player's demographics (age, educational background, etc.) influence the quality and type of knowledge elicited by LLMs in GWAPs?

Player demographics also influence the quality and type of knowledge elicited by LLMs in GWAPs. Studies show that factors such as age, educational background, and cultural context impact how users interact with GWAPs and the knowledge they contribute. Younger players and those with higher educational backgrounds often provide more diverse and complex semantic associations due to their broader vocabulary and cognitive skills [3].

Cultural context also plays a crucial role, with players from different backgrounds offering unique insights that enrich the dataset [12]. Age and education intersect to affect interaction styles and engagement levels; older adults may approach tasks more cautiously, while younger individuals might show higher creativity and risk-taking. Educational background influences the knowledge base and critical thinking approaches of players, essential for complex knowledge elicitation.

Age-related differences in cognitive processing also show that variations in task difficulty and processing speed can significantly impact the quality of knowledge elicited [10]. For instance, older adults may require more time to process and respond to tasks compared to younger individuals, which can influence the complexity and detail of the knowledge they provide.

Moreover, demographic factors also influence the dynamics of player collaboration and competition in GWAPs, which can affect the knowledge elicitation process. As an example, players from different age groups and educational backgrounds may have varying preferences for collaborative versus competitive game mechanics. Younger players might prefer competitive elements that drive rapid and diverse responses, while older players might favor collaborative approaches that allow for more thoughtful and detailed contributions. Educational background can similarly impact these preferences, with more academically inclined individuals possibly favoring tasks that require deeper analytical thinking and cooperation [21].

5.2 Synthesis

By examining the three sub-questions, we derive a comprehensive understanding of what types of knowledge LLMs in GWAPs can elicit. The main research question ”**What type of knowledge can be elicited using LLMs in GWAPs?**” will be answered by synthesising the findings from the sub-questions and other findings from academic sources.

Our findings indicate that LLMs have substantial potential to enhance the knowledge elicitation capabilities of GWAPs. The integration of LLMs in GWAPs can elicit various types of knowledge, including semantic associations, factual information, commonsense knowledge, contextual insights, and experiential knowledge.

Firstly, we see that LLMs are great in capturing complex semantic associations. This is reflecting how users perceive and relate different concepts. This type of knowledge is crucial for applications in language processing and AI training. Games like Word Ladders demonstrate that LLMs can engage users dynamically, uncovering associations and contextually specific meanings that traditional methods might overlook. This ability to find semantic connections can significantly enhance the worthiness of data collected through GWAPs.

Secondly, LLMs can effectively elicit factual information and especially when integrated into educational games. This includes detailed information and accurate data on a wide range of subjects. The capability of LLMs to provide precise responses makes them well-suited for eliciting structured factual knowledge in quiz-based and educational contexts. This type of knowledge transfer not only enhances user learning but also ensures the retention of accurate information through engaging and interactive methods.

Thirdly, commonsense knowledge which includes everyday practical knowledge and social norms can be gathered through scenarios that require players to apply their understanding of the world around them [6]. LLMs can facilitate games that simulate real-life situations, prompting users to provide insights based on their commonsense reasoning. This type of knowledge is valuable for training AI systems to better understand human behavior and decision-making processes.

Additionally, our findings have shown that contextual insights are another critical type of knowledge that can be elicited through LLM-enhanced GWAPs. Contextual insights involve understanding the situational and background factors of the players that influence how knowledge is applied. By engaging users in scenarios requiring context-specific decision-making, LLMs can help capture the nuances of how context affects knowledge application. This is particularly useful in personalized learning environments and adaptive interfaces where context plays a pivotal role.

Moreover, they can elicit experiential knowledge, such as personal insights and subjective understanding that users hold about the game. This type of knowledge encapsulates individual opinions and understanding concerning diverse topics. For games that prompt user-based personal experiences and reflections, LLM can capture such rich qualitative data. This might be overlooked by more traditional methods of data collection. This approach enables the collecting of different views and a richer context concerning information for the overall dataset, thereby increasing its quality by valuable qualitative insights.

In conclusion, we have found that integrating LLMs into GWAPs presents a promising approach to eliciting a wide range of knowledge types. Semantic associations, factual information, commonsense knowledge, contextual insights, and experiential knowledge can all be effectively elicited through these GWAPs. By leveraging the strengths of both LLMs and GWAPs, and understanding the demographic factors that influence player contributions, it is possible to create more engaging and effective platforms for data collection and knowledge discovery. Future research should continue to explore innovative ways to span the diverse contributions of players to enhance the quality and depth of knowledge elicited in GWAPs.

6 Discussion

In this section, we will discuss the findings from the reviewed literature, place them in a broader context, and reflect on the overall conclusions. The goal is to integrate insights from various studies and discuss their implications.

6.1 Implications

The findings from this study have several important implications for both theoretical and practical applications. The integration of LLMs into GWAPs show quite a significant advancement in artificial intelligence and gamification. Theoretically speaking, our findings support the hypothesis that LLMs can enhance the types of knowledge elicited through GWAPs by providing more engaging and contextually relevant interactions for users. This integration allows for the elicitation of multiple types of knowledge, including semantic associations, factual information, and user preferences, which would be more difficult to elicit through traditional methods. This aligns with previous research indicating the potential of LLMs to generate human-like responses that can be used in different applications, which are including academic research and practical applications in different industries [16; 8].

The use of LLMs in GWAPs can change the tedious data annotation tasks into some interactive and enjoyable activities. This is then attracting a larger pool of participants and improving the quality and quantity of the data collected. This double benefit of entertainment and utility, as demonstrated by projects like reCAPTCHA, highlights the potential for more application in different fields such as education, data annotation, and even citizen science [23]. Moreover, the findings, as a result of the different types of knowledge that can be elicited, suggest that the integration of LLMs can facilitate the exploration of complex semantic relationships and improve the learning experience in educational games [3].

These contributions challenge existing theories by demonstrating the effectiveness of combining LLMs and GWAPs to enhance user engagement and knowledge elicitation. We think that the implications will extend to influencing practice, policy, and further research in the field. Moreover, suggesting new directions for leveraging AI in gamified environments to achieve educational and data collection goals.

6.2 Research Process

When reflecting on the process of conducting the literature review, the PRISMA workflow proved to be an effective methodology for systematically collecting and analyzing the relevant literature within the field of our research. The structured approach ensured transparency and reproducibility, which are critical for the integrity of any literature survey [15]. The initial identification of studies through academic databases like Google Scholar which was followed by fully screening and eligibility checks. This allowed us to narrow down to a highly relevant set of papers for our research purposes.

A comprehensive overview of the current state of research on the integration of LLMs into GWAPs was provided by the synthesis of these studies, where the types of knowledge that can be elicited were extracted. The methodologies used in the reviewed papers varied. The papers has included methods from investigations to theoretical analyses. This is offering insights from multiple perspectives into the capabilities and limitations of LLMs in a gamified environment. This approach enabled us to identify key trends, gaps, and emerging themes. This provided us a robust foundation for addressing our research questions.

6.3 Impact on Society

The impact of integrating LLMs into GWAPs goes beyond this academic research to societal implications. The use of gamified environments for data collection and knowledge elicitation can give access to scientific research. This allows individuals from diverse backgrounds to contribute valuable insights, even without knowing the full potentials. This approach can enhance public engagement with science and technology and creating a better understanding and appreciation of the capabilities of AI.

However, we cannot ignore the ethical dilemmas associated with AI and gamification. Issues like data privacy, obtaining user consent, and the potential for bias in AI-generated responses must be carefully addressed. It is crucial to manage these aspects to ensure that the benefits of these technologies are enjoyed without compromising ethical standards. Future research should focus on developing frameworks and guidelines to responsibly use LLMs in GWAPs, ensuring that they are both effective and ethically correct.

6.4 Future Work

Our research, therefore, involves pointing out some other open issues, and more studies are needed. First, a need exists for more empirical studies to assess the long-term impact of LLMs on the user engagement and learning outcomes in GWAPs. Although our findings do show that LLMs might have the capacity to enhance knowledge elicitation in its users, it is vital to evaluate their effectiveness over extended periods of time using long-term studies.

It is also essential to develop more filtering and validation techniques to improve the quality of the data that LLMs will have collected. Such a method of filtering the noise in the data while improving the accuracy and relevance of the knowledge gathered, would significantly improve the utility of LLM-enhanced GWAPs.

Moreover, it is necessary to elaborate on the fact that player demographics might have a significant impact on the level of success of LLMs in GWAPs. What this does, is that it helps in understanding how such factors as age, educational background, and cultural context will affect the interaction of end-users, thereby leading to gamified environments. It is only through such dedicated collaborations among AI researchers, educators, game designers, and ethicists that the field at large can realize such advances drawn from diversified perspectives and expertises. We should work towards a more robust and ethical frameworks to integrate LLMs into GWAPs, thereby improving their impact on education, data collection, and public engagement with science.

One promising idea is to do a long-term study of user interaction conducted with LLM-enhanced GWAPs over a few months. Thus, to realize how different demographic backgrounds influence participants' behavior. We have to take some factors in considerations, such as age, educational background, and cultural context can be taken into account. The study can be implemented with pre- and post-interaction surveys where users' knowledge and engagement levels are measured. In addition, the existing mechanism of data collection may work on strengthening the filtering and validation process, which can be still optimized for the effectiveness in noise reduction for further quality augmentation of the collected knowledge.

In conclusion, our primary research was understanding how LLMs can be integrated into GWAPs for knowledge elicitation and to find out which types of knowledge can be elicited from the GWAPs. Working in this way, the challenges identified may be overcome and the strengths of the two technologies are exploited to bring forth more successful and engaging platforms for data collection and knowledge discovery. Insights from this research can be helpful for future research and development in this promising area.

7 Responsible Research

In addressing the aspects of responsible research for this project, it is essential to highlight the measures taken to ensure both ethical integrity and reproducibility.

7.1 Reproducibility

The reproducibility of our methodology is a critical aspect of this research and to ensure the reproducibility of this literature survey, a systematic approach was used according to the PRISMA framework. The methodology section of this paper explained how the framework was used and what query was used. This transparency allows other researchers to replicate the study, verify the results, and build upon the findings presented here.

All sources and references are documented, enabling others to access the same materials and replicate the review process. By providing clear citations and detailed descriptions of the research process, this study ensures that its methods can be independently verified and validated.

7.2 Ethical Considerations

This research adheres strictly to ethical standards by using only publicly available and ethically sourced literature. The reviewed articles are accessed through academic databases, such as Google Scholar, ensuring that all data is legally and ethically obtained.

Moreover, the study maintains the confidentiality and integrity of the sources by accurately citing all references and avoiding any form of plagiarism. The selection of literature is conducted with openness, ensuring a balanced representation of various perspectives and avoiding any potential biases.

In addition, the literature survey is conducted in a manner that respects the intellectual property rights of the original authors. All findings are presented with proper attribution, and no proprietary data or sensitive information is used without appropriate permissions.

8 Limitations

In this section, we will evaluate the diversity and representativeness of the sources reviewed, and address any potential biases in the selection of literature and how they have been mitigated.

8.1 Diversity and Representativeness of Sources

Evaluating the diversity and representativeness of the sources reviewed is essential to ensure a valid and balanced perspective in our survey.

Our literature survey includes a wide range of sources from academic databases, such as Google Scholar. These sources give access to academic papers, which include theoretical analyses and practical applications of LLMs in GWAPs, providing a great overview of the field.

While we have strived to include diverse perspectives, it is possible that some viewpoints may be underrepresented. The majority of the reviewed papers originate from academic sources from technologically advanced regions, which may not fully capture the global landscape of AI, LLMs and gamification research. Future surveys should aim to include more research from underrepresented regions to provide a more complete survey.

8.2 Potential Biases and Mitigation

Acknowledging and addressing potential biases in the selection of literature is crucial for maintaining the integrity of our research.

The inclusion of studies was based on predefined criteria to minimize selection bias. However, there is always a risk of excluding relevant papers that do not fit the initial criteria. To mitigate this, multiple rounds of screening and review were conducted, looking at the papers from different perspectives, mostly inspired by the different sub-questions to create those perspectives.

Positive results are more likely to be published than negative or inconclusive findings, which can then give a gap in the overall conclusions of a literature review. We have attempted to mitigate this by including preprints, which may not have undergone the same publication checks but provide valuable insights into ongoing research.

In conclusion, while our study has made significant effort in understanding the integration of LLMs into GWAPs, it is important to recognize the limitations and areas for improvement. By addressing ethical considerations, ensuring reproducibility, enhancing the diversity of sources, and mitigating potential biases, future research can build on our findings to further advance this promising field of knowledge elicitation.

9 Conclusions

This research investigates the integration of Large Language Models (LLMs) into Games With A Purpose (GWAPs) to determine the types of knowledge that can be elicited through this combination. The primary research question, **”What type of knowledge can be elicited using LLMs in GWAPs?”**, focuses on identifying the potential of LLMs to enhance the effectiveness of GWAPs in collecting diverse types of knowledge. The The sub-questions explored include:

1. What kinds of semantic associations do players reveal when interacting with LLMs in GWAPs?
2. What are the challenges and limitations of using LLMs to elicit specific types of knowledge in a GWAP setting?
3. How does the player's demographics influence the quality and type of knowledge elicited by LLMs in GWAPs?

The integration of LLMs into GWAPs enables the elicitation of different knowledge types, enhancing both the quantity and quality of data collected in gamified environments. The primary types of knowledge identified include semantic associations, factual information, commonsense knowledge, contextual insights, and experiential knowledge.

One of the significant strengths of LLMs is the ability to capture complex semantic relationships. Live interaction with players shows how the more traditional methods can easily overlook connections between words and concepts. For example in the application Word Ladders, players will create semantic associations between words for the relations and meanings bound to the different contexts. This knowledge is critical to applications in language processing and the training of artificial intelligence because it provides data on how end-users perceive and relate to different concepts.

LLMs can support the elicitation of factual knowledge well and particularly in the educational games. Due to their great knowledge, they can provide factually accurate answers, fitting well for quiz-based games and other educational applications. This helps users improve their understanding of facts in a fun manner.

LLMs allow the effective collection of commonsense knowledge derived from the everyday use of practical knowledge and social norms. Development of the games mirroring real-life situations, the user can bring his understanding of the world into play. Therefore it is capturing insights based on the commonsense reasoning explained. While such knowledge is valuable for training AI systems better to understand the behavior of humans and processes of decision-making.

Contextual insights involve understanding the situational and background factors that influence how knowledge is applied. LLM enhanced GWAPs can engage users in scenarios requiring context specific decision-making, capturing the nuances of how context affects knowledge application. This is particularly useful in personalized learning environments and adaptive interfaces where context plays a pivotal role.

Experiential knowledge is the personal experience and insight that the users bring to the game. This is usually highly subjective knowledge and uniquely offered according to different themes. By encouraging users to share their experiences in the design of the game, LLMs help collect qualitative data that might be easily missed with structured techniques of data collection.

The results show that LLMs can help GWAPs in more interactive and effective knowledge-elicitation tools. One of the most significant of these contributions is the advancement of user involvement through the enhancement of interest that LLMs provide hence attracting more varied participants and enhancing the quality of the data collected. The elicitation procedures of LLM-enhanced GWAPs are not limited to semantic associations and factual information. They come from common sense, context, and experiential knowledge.

However, integration of LLMs into games brings challenges regarding data noise, user engagement, computational resources, and ethical considerations, such as bias. This research addresses with insights and ideas to mitigate those issues. It is important for this research to provide insights and strategies that will ensure the benefits of these technologies without compromising the ethical standards.

In conclusion, a wide variety of knowledge types elicited from GWAPs using the integration of them with LLMs. Leveraging the strengths of both LLMs and GWAPs and understanding the demographic factors that influence player contributions can lead to more engaging and efficacious platforms for data collection and knowledge discovery. This highlights the need for continued research on innovative ways to combine LLMs and GWAPs for further development of AI technologies and data collection to be of significant interest for future research and development in this promising area.

References

[1] 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*, 2022.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[3] Marianna Marcella Bolognesi, Claudia Collacciani, Andrea Ferrari, Francesca Genovese, Tommaso Lamarra, Adele Loia, Giulia Rambelli, Andrea Amelio Ravelli, and Caterina Villani. Word ladders: A mobile application for semantic data collection, 2024.

[4] Rishi Bommasani et al. On the opportunities and risks of foundation models. *Stanford University*, 2021.

[5] Tom B. Brown et al. Language models are few-shot learners, 2020.

[6] Erik Cambria, Dheeraj Rajagopal, Kenneth Kwok, and Jose Sepulveda. Game engine for commonsense knowledge acquisition. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2022.

[7] Wenhu Chen. Large language models are few(1)-shot table reasoners, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

[10] Kelly C. Harris, Mark A. Eckert, Jayne B. Ahlstrom, and Judy R. Dubno. Age-related differences in gap detection: Effects of task difficulty and cognitive ability. *Hearing Research*, 264(1):21–29, 2010.

[11] Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. Exploring human-like translation strategy with large language models, 2023.

[12] Abdelraouf Hecham, Madalina Croitoru, Pierre Bisquert, and Patrice Buche. Extending gwaps for building profile aware associative networks. In *Graph-Based Representation and Reasoning*. Springer International Publishing, 2016.

[13] Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Hassan Foroosh, Dong Yu, and Fei Liu. Sportsmetrics: Blending text and numerical data to understand information fusion in llms, 2024.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[15] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLOS Medicine*, 2009.

[16] Alec Radford et al. Improving language understanding by generative pre-training. *OpenAI Blog*, 2018.

[17] Mark O. Riedl and Alexander Zook. Ai for game production. In *IEEE Conference on Games (CoG)*. IEEE, 2019.

[18] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[19] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts, 2020.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[21] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[22] Luis von Ahn and Laura A. Dabbish. Labeling images with a computer game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004.

[23] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. Recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[24] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. From human-computer interaction to human-ai interaction: new challenges and opportunities for enabling human-centered ai, 2021.