From Text to Text Game: A Novel RAG Approach to Gamifying Anthropological Literature and Build Thick Games

Michael Hoffmann¹[®]^a, Jan Fillies^{1,2}[®]^b, Silvio Peikert³[®]^c and Adrian Paschke^{1,2,3}[®]^d

¹Freie Universität Berlin, Berlin, Germany
²Institut für Angewandte Informatik, Leipzig, Germany
³Fraunhofer-Institut für Offene Kommunikationssysteme, Berlin, Germany

- Keywords: Text Games, Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Game-Based Learning, AI in Education, Computational Anthropology.
- Abstract: This study introduces a novel approach to gamifying anthropological literature using Large Language Models (LLM), specifically GPT-3.5, to create text-based games. Traditional methods of gamifying specialized literature often require costly game design and programming expertise. The method proposed by the authors employs Retrieval Augmented Generation (RAG) to transform anthropological classics into interactive games, potentially expanding the audience for anthropological knowledge. To evaluate this prototype, the researchers developed a corpus of 50 university-level exam questions with human-annotated gold standard answers. Together with an expert in social anthropology, they compared RAG-generated responses to these questions against both the gold standard and non-RAG approaches using a self-designed metric called ATGE (Anthropological-Text-Game Evaluation), which assesses the general quality, ethnographic depth, and honesty of the answer. Results indicate that the RAG-based system outperforms a non-RAG approach in factual accuracy and retention of ethnographic details, though it remains inferior to human-annotated answers. This suggests that RAG-based gamification can create 'thick games' with substantial ethnographic depth, offering a promising, cost-effective method for making anthropological insights more accessible in an educational setting while maintaining scholarly integrity.

1 INTRODUCTION

In recent years, the intersection of technology and education has opened up new avenues for knowledge dissemination and engagement. One particularly promising area is the gamification of academic content, which has shown potential in increasing student engagement (Bouchrika et al., 2021; Wang and Tahir, 2020) and retention of complex concepts (van Horssen et al., 2023; Rizki et al., 2024). However, the traditional approach to creating educational games often requires significant investment in game design and programming expertise, making it a costly endeavor for many academic disciplines (Freire et al., 2023). This paper introduces an innovative and cost-effective method for gamifying anthropological literature using Large Language Models (LLMs), specifically GPT-3.5, to create interactive text-based games.

- ^a https://orcid.org/0000-0001-5003-5138
- ^b https://orcid.org/0000-0002-2997-4656
- ^c https://orcid.org/0000-0001-5716-1540
- ^d https://orcid.org/0000-0003-3156-9040

At the core of this study's approach is the use of Retrieval Augmented Generation (RAG) (Lewis et al., 2020), a technique that combines the vast knowledge base of LLMs with specific, curated information from text. This study applies a Retrieval Augmented Generation based approach to transform an anthropological classic text into a text-based game, building upon existing research (Pia, 2019; Hoffmann et al., 2024). While previous research highlighted the demand for anthropological games among researchers (Pia, 2019) and demonstrated the potential of non-RAG Large Language Models in creating meaningful games based on internet resources (Hoffmann et al., 2024), the current study expands this work through three distinct contributions:

• 1. Development of a RAG-based prototype system: This system enables the conversion of an anthropological monograph into an interactive textbased game, securing that the LLM has the exact right knowledge, thereby potentially offering richer content and more accurate representations of the source material than with a simple non-RAG approach.

246

Hoffmann, M., Fillies, J., Peikert, S. and Paschke, A. From Text to Text Game: A Novel RAG Approach to Gamifying Anthropological Literature and Build Thick Games. DOI: 10.5220/0013215400003932 Paper published under CC license (CC BY-NC-ND 4.0) In Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2, pages 246-256 ISBN: 978-989-758-746-7; ISSN: 2184-5026 Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

- 2. Creation of a benchmark dataset: The authors compile a corpus of 50 classroom questions from various sources, establishing a Q&A benchmark dataset to assess RAG-based and Non-RAG based Large Language Models' capacity to address academic inquiries in anthropology based on one specific piece of literature.
- 3. Comprehensive evaluation: The study evaluates the prototype using the benchmark dataset, employing two different models and a quantitative semantic similarity comparison. Additionally, two trained annotators of the research team, together with a domain expert, manually compare the models' outcomes with the benchmark dataset, providing a multi-faceted assessment.

The results reveal that the RAG-based system outperforms its non-RAG counterpart in factual accuracy, ethnographic depth, and honesty though it still falls short to the quality of the human-annotated goldstandard answers. Nevertheless, this finding suggests that RAG-based approaches hold promise for creating what the authors call 'thick games': interactive experiences that encapsulate significant ethnographic depth while maintaining scholarly integrity. By leveraging RAG technology, this study paves the way for more accessible and engaging anthropological content, potentially bridging the gap between academic research and public engagement in the field.

In the following, this research reviews relevant literature, details the prototype's design including data inputs, gamification elements, prompt engineering, and technical architecture. The evaluation section presents the dataset, discusses automatic and manual assessment methods, and analyzes results. The study concludes with findings and future research directions.

2 RELEVANT LITERATURE

2.1 LLMs and Retrieval Augmented Generation

Since the unveiling of the transformer neural network model in the paper "Attention Is All You Need" by Google Brain researchers in 2017 (Vaswani, 2017), Large Language Models like GPT-3.5/4, LLama (Touvron et al., 2023), and Palm (Chowdhery et al., 2023) have demonstrated remarkable capabilities in tasks such as text generation, complex questionanswering, and information retrieval. Recent advancements have further improved the performance of language models by incorporating vast amounts of retrieval data. Borgeaud et al. (Borgeaud et al., 2022) demonstrated significant improvements in language model performance by retrieving from trillions of tokens, highlighting the potential of retrieval-based approaches in enhancing LLM capabilities. Therefore, it is evident that LLMs can acquire extensive in-depth knowledge from data and serve as parameterized knowledge bases (Petroni et al., 2019). However, researchers have identified various limitations in LLMs in recent years, including their inability to expand, revise, or update their memory, their lack of explainability in predictions, and susceptibility to hallucinations (Marcus, 2020).

A significant part of research on LLMs has extensively concentrated on enhancing their performance, leading to the emergence of two primary strategies. Firstly, LLMs can undergo optimization through finetuning, a process of 'rebaking' specific weights within the initial parameterized knowledge base. In contrast, the Retrieval Augmentation Generation approach, introduced by Lewis et al. (Lewis et al., 2020), aims to address LLM performance issues by incorporating a non-parametric knowledge base alongside the originally trained LLM. This incorporation of a second non-parameterized knowledge base takes the form of a vector base and offers two key advantages: the ability to swiftly include unstructured data by converting it into a vector base and the flexibility to update and revise knowledge within the vector base (Manathunga and Illangasekara, 2023). This study introduces a novel application of Retrieval Augmented Generation to the domain of anthropological games. More specifically, the prototype developed in this study employs what Gao et al (Gao et al., 2023) have termed the naive RAG approach, following a traditional process of indexing, retrieval, and generation, but applied for the first time to create interactive experiences based on anthropological texts.

The naive RAG (Gao et al., 2023) process begins with the cleansing and extraction of data from the original anthropological text, followed by 'chunking' it into smaller, manageable pieces. These chunks are transformed into vector representations and stored in a database for efficient retrieval. When a user interacts with the anthropological game, their input is converted into a vector representation, and the system retrieves the most relevant text chunks based on similarity scores. While the naive RAG approach has known limitations in retrieval and generation quality (Gao et al., 2023), its application to anthropological games represents a novel contribution to the emerging corpus of literature on Retrieval Augmented Generation.

2.2 Evaluation of RAG-Based LLMs

Previous research has developed various methods for evaluating RAG-based LLMs, which can be broadly categorized into two main approaches. The first involves evaluating a RAG-based system by comparing its output to a ground truth using human evaluators. This method allows for the creation of custom metrics that capture the subtleties and nuances of the generated outputs when compared to a benchmark dataset. However, it is expensive and introduces the risk of bias, as humans are used as judges for the system's performance.

The second category centers on automated evaluation methods, primarily using established NLP metrics to assess RAG performance. These include BERTScore (Zhang et al., 2019), which evaluates generated outputs by computing cosine similarities between token embeddings of generated and reference answers. Other researchers employ metrics like Rouge-1 and SemScore (Aynetdinov and Akbik, 2024) for answer generation evaluation, while metrics such as HIT Rate, Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) or Mean Reciprocal Rank (MRR) are used to assess retrieval quality (Liu, 2023). However, these approaches all require human-annotated reference data.

In cases where human annotations are unavailable, automated frameworks like RAGAS(Es et al., 2023) have been proposed. RAGAS evaluates the performance of RAG-based systems by assessing context relevance for the retriever and answer relevance for the generator, with an additional metric called faithfulness to measure the coherence between the two modules. Similarly, ARES(Saad-Falcon et al., 2023) follows a similar approach, evaluating RAG-based systems across these same dimensions. In this work, the authors employ both human and automated evaluation methods to assess their RAG-based prototype. By integrating these approaches, the authors offer a unique contribution to the literature, where an entire monograph from a subfield of the humanities is used as context for the RAG-based system's qualitative and quantitative evaluation.

2.3 Educational Potential of Digital Games for the Humanities

The integration of digital games into educational settings has emerged as a significant area of research in recent decades. This has sparked debate around the effectiveness of game-based learning in academic settings (Ashinoff, 2014). While media time and again presented skeptical findings (Ashinoff, 2014,

p. 1109),(Huesmann, 2010) it is important to remember that James Paul Gee argued already in 2003 that well-designed video games are essentially learning machines - and therefore well equipped for effective learning of academic content - due to their ability to motivate and engage players (Gee, 2003). Recent research supports this claim, revealing the potential of digital games as pedagogical tools (Wang and Tahir, 2020; van Horssen et al., 2023). Wang et al (Wang and Tahir, 2020), for example, demonstrated that Kahoot! has positive effects on learning performance, classroom dynamics, attitudes, and anxiety. Similarly, Van Horssen et al's research on commercial strategy games like Sid Meier's Civilization IV: Colonization (Firaxis Games 2008) and GreedFall (Spiders 2019) in higher education humanities curricula showed that students demonstrated enhanced engagement with historical narratives and developed stronger critical thinking skills through holistic reflexive engagement (van Horssen et al., 2023).

However, this scholarly interest in digital tools remains largely limited to video games, while textbased games remain understudied as potential learning tools despite their rich fifty-year history (Reed, 2023). This oversight is notable since text-based games' focus on written narrative makes them particularly well-suited for humanities education - their format naturally emphasizes close reading, contextual analysis, and meaning-making without visual distractions. Recent examples demonstrate this educational potential. For example, Andrea Pia's game 'The Long Day of Young Peng'¹ uses interactive narrative to illuminate the complexities of rural-to-urban migration in contemporary China, conveying nuanced social dynamics and personal experiences more effectively than traditional teaching materials(Pia, 2019). Similarly, Manuba (Manuaba, 2017) leverages Indonesian folklore in a text-based game adaptation of "Danau Toba" designed to enhance players' reading comprehension.

Building upon these insights, the current study proposes to address this research gap by implementing a Retrieval-Augmented Generation approach to convert anthropological text into a text-based game. This methodology aims to enhance both the accuracy and ethnographic depth of generated content while preserving the engaging, interactive characteristics that make text-based games effective educational tools. The aim is to create "thick" gaming experiences that immerse students in rich, contextual understanding of cultural phenomena, social relationships, and lived experiences. Just as anthropologists strive to create "thick descriptions"(Geertz,

¹http://thelongdayofyoungpeng.com

2008) in their ethnographic work, the authors of this study claim that educational games in anthropology should aim to be "thick games". That is to say that the goal is to create educational games that not only effectively convey anthropological knowledge but also maintain the methodological rigor and depth characteristic of anthropological research. By approximating this goal, this study contributes to one of the key challenges in educational game design for the humanities: maintaining academic rigor while ensuring student engagement.

2.4 Automatic Game Generation

Within the broad field of game AI, research into automatic game generation traces its origins to the early 1990s, beginning with METAGAMES by Pell (Pell, 1992) in 1992. This pioneering work laid the foundation for future developments, including significant advances through evolutionary algorithms, as demonstrated by Cameron Browne's 2008(Browne, 2008) research that produced playable games like Yavalath².

The introduction of Large Language Models (LLMs) has fundamentally transformed automatic game generation, offering new possibilities for creating dynamic, responsive gaming experiences (Gallotta et al., 2024). Researchers have combined LLMs with evolutionary algorithms to generate and optimize game concepts (Todd et al., 2024), while commercial applications like AI Dungeon³ have demonstrated the potential for creating personalized, open-ended narratives using GPT-3 (Hua and Raley, 2020).

In the context of text-based anthropological games, recent work by Hoffmann et al. (Hoffmann et al., 2024) has explored using GPT-3.5 to develop anthropological text-based games, creating four distinct designs but also revealing limitations in narrative coherence and gameplay innovation. This study builds upon such prior works and extends them through RAG-based approaches. It therefore contributes to the literature of automatic game generation by presenting a use-case for the creation of a RAG-based text-game based on a singular text in the domain of anthropology.

3 PROTOTYPE DESIGN

This section presents the design and architecture of the developed RAG-based prototype. It begins by discussing the authors' rationale for selecting the anthropologist Bronislaw Malinowski's book 'Argonauts of the Western Pacific' (Malinowski, 2013) as the central text for gamifying with the RAG-based application. The section then outlines the type of game the prototype will generate, followed by a detailed exploration of the system prompt used as well as the prototype's technical design and implementation.

3.1 Anthropological Monograph as Input Data

For this study, the authors selected the anthropological classic 'Argonauts of the Western Pacific' (1922) by Bronislaw Malinowski (Malinowski, 2013) as the primary text to gamify using the prototype. This seminal work was chosen for several compelling reasons. First, the book is predominantly text-based with minimal images, making it ideal for text-centric extraction and processing. This characteristic aligns well with the capabilities of Large Language Models, which excel at understanding and generating textual content. The dense, descriptive nature of Malinowski's writing provides a rich foundation for the LLM to draw upon, ensuring that the generated game content is substantive and true to the original ethnographic observations.

Secondly, 'Argonauts of the Western Pacific' is renowned for its ethnographic depth, a quality that has been consistently praised by reviewers and scholars in the field of anthropology(Leach and Leach, 1983; Hann and James, 2024). This depth of cultural description and analysis makes the book an excellent candidate for transformation into an interactive, educational game. The details of Trobriand Islander life, the complex Kula⁴ exchange system, and Malinowski's reflections on the practice of ethnography itself offer a wealth of material for creating engaging scenarios, thought-provoking questions, and immersive role-playing experiences. By using such a rich text, the authors aim to demonstrate the potential of their RAG-based system to capture and convey nuanced anthropological concepts in an interactive format.

Lastly, the choice of this particular monograph serves a broader purpose in the context of anthropological education. 'Argonauts of the Western Pacific' is not only a foundational text in the field but also one that continues to be widely read and discussed in anthropology courses worldwide (Hann and James,

²Originally created by an AI named LUDI, Yavalath stands as an abstract strategy game playable by groups of two or three (See http://cambolbro.com/games/yavalath/)

³https://aidungeon.com

⁴The Kula Ring is an elaborate inter-island exchange system involving communities in the Trobriand Islands and surrounding archipelagos. It involves the ceremonial circulation of two types of valuables governed by strict rules.

2024). By creating a game based on this text, the authors seek to provide a novel, interactive method for students to engage with classic anthropological literature. This approach has the potential to make the sometimes challenging content more accessible and engaging to modern students, while also preserving the intellectual rigor and ethnographic insight of the original work.

3.2 Gamified Approach

The authors developed 'Topic Quiz', a RAG-based text game that builds on the game design framework established by Hoffmann et al. (Hoffmann et al., 2024). This design was specifically chosen as it facilitates to evaluate LLM outputs, particularly their factual accuracy and cultural authenticity when compared to primary source materials. The game operates as follows:

Phase 1 - Theme Selection. In a first phase, a university teacher - or seminar leader - identifies three central topics from 'Argonauts of the Western Pacific'. For example, available themes may span key anthropological concepts such as the Kula ring system, indigenous farming methods and community organization, magical beliefs and practices in Trobriand culture, or reflections on Malinowski's own positionality within a broader geopolitical perspective.

Phase 2 - AI Dialogue. In a second phase, the players (students) pick one theme card and participate in a focused 10-minute conversation with an AI system embodying Bronislaw Malinowski. This interactive component allows for deeper comprehension of the ethnographic content and Malinowski's experiences through direct exploration of the author's viewpoint.

Phase 3 - Knowledge Assessment. In a third phase, the players are presented with and have to answer a set of ten teacher-crafted multiple choice questions that evaluate their grasp of both the source text and insights derived from the AI interaction. Individual scores are then recorded and made public on a shared leaderboard. The latter introduces a competitive aspect designed to spark ongoing participation and multiple gameplay attempts.

3.3 System Prompt Design for Gamification

After extensive experimentation with different system prompt designs over a period of one week, the authors developed a specialized prompt that enables the LLM to effectively embody the role of the anthropological book's author, Bronislaw Malinowski, and generate answers of appropriate length and format. The final system prompt reads as follows:

System Prompt. You are the renowned anthropologist Bronislaw Malinowski. You are known for your work in ethnography and your development of functionalism in anthropology. Respond to questions about the document as Malinowski would, drawing on your expertise in anthropology and your fieldwork experiences. Use your knowledge to provide insightful analysis and commentary and limit your answers to a maximum of three paragraphs and 200 words.

Context: context

Human: question Bronislaw Malinowski:

3.4 Technical Design and Implementation of the RAG Prototype

The implemented prototype follows a classic Retrieval-Augmented Generation architecture, designed to process and interact with anthropological texts in an intelligent manner. The system operates through several sequential stages, beginning with the ingestion of unstructured text from anthropological monographs and articles. Using recursive character text splitting, the content is segmented into manageable chunks, which are then converted into vector embeddings. After evaluating different vector storage solutions, the team implemented Facebook AI Similarity Search (FAISS) over ChromaDB⁵ for the vector database.

The system's core functionality relies on a specialized retriever that searches the vector database when users submit queries. This retriever is designed to constrain the language model's responses to the available local context in the vector store, implemented through a template-guided response generation. To enhance user interaction, the system implements a conversational retrieval chain that maintains context across multiple interactions, enabling more coherent and contextually aware dialogues.

The prototype was developed as a web application with a Streamlit⁶ frontend and a Python backend using the Langchain⁷ library. The source code is publicly available⁸, though users must provide their own OpenAI API key. The interface enables users to upload anthropological texts in PDF format and engage in a 'Topic Quiz' game - a topic-specific discussion

⁵https://pypi.org/project/chromadb/

⁶https://streamlit.io

⁷https://www.langchain.com

⁸The source code is accessible at: https://github.com/michaelpeterhoffmann/ragAnthro

with an AI simulation of the text's author. In the interface, users are represented by a red anthropomorphic icon, while Malinowski's simulation uses a yellow robot icon. This deliberate choice of a robot icon clearly signals to users that they are interacting with an AI rather than a human (See Figure 1).



Figure 1: A Screenshot of the Prototype Game Topic Quiz.

The backend architecture leverages Langchain's capabilities to create a sophisticated interaction system. It employs a ChatPromptTemplate that combines context about Bronislaw Malinowski's background and expertise with user questions. Through a custom SystemMessagePromptTemplate, the language model is instructed to respond in Malinowski's persona, while the ConversationalRetrievalChain ensures coherent dialogue maintenance. To ensure reproducible results, the system uses OpenAI's GPT-3.5-turbo model with the temperature parameter set to 0.

4 EVALUATION

The systematic evaluation of generative Large Language Model outputs presents significant methodological challenges stemming from the nondeterministic nature of natural language generation tasks. The fundamental complexity arises from the inherent variability in human-generated responses, where a single input query can lead to multiple valid outputs, making the definition of "correct" and "appropriate" responses computationally non-trivial.

Contemporary evaluation methodologies for Retrieval-Augmented Generation systems typically leverage standardized benchmark datasets that provide reference outputs for validation purposes. These datasets serve as established ground truth corpora against which model performance can be systematically assessed. However, in specialized domains such as anthropological text analysis, specifically for Question-Answering (QA) systems, there exists to the best of the authors knowledge a significant gap in the availability of domain-specific evaluation datasets.

To address this limitation in the context of anthropological QA evaluation, the research team, in collaboration with expert senior anthropologists, constructed a specialized evaluation corpus comprising 50 question-answer pairs derived from the anthropological monograph 'Argonauts of the Western Pacific'. This carefully curated dataset serves as the domain-specific ground truth for the evaluation process, which was undertaken in two ways. On the hand through automated evaluation utilizing established computational metrics; on the other hand through manual assessment implementing a novel, purposebuilt and anthropology-centered evaluation metric.

The subsequent sections lay out the evaluation analysis in three parts: first, the composition and characteristics of the evaluation dataset; second, the implementation of multiple automated evaluation metrics and their results; and third, the development and application of a custom evaluation protocol for manual assessment.

4.1 The Evaluation Dataset

The evaluation dataset consists of 50 manually curated question-answer pairs, specifically designed for the Topic Quiz Game implementation. The dataset construction process prioritized question diversity to simulate realistic queries from the target audience, including students and academic researchers. Following the taxonomic framework proposed by Cherumanal et al (Cherumanal et al., 2024, p. 2), the questions were systematically categorized into three distinct types:

- Out-of-Scope Questions: Queries intentionally constructed without definitive answers due to content absence in the source material.
- Simple Questions: Queries with answers locatable within single text passages.
- Complex Questions: Queries requiring information synthesis across multiple passages, demanding comprehensive content understanding.

Table 3 (See Appendix) provides representative examples from each category of the constructed dataset.

4.2 Automatic Evaluation

4.2.1 Metrics Used

The authors employed several established metrics to evaluate different aspects of the prototype system:

- NDCG (Normalized Discounted Cumulative Gain): Used to assess retrieval effectiveness. NDCG measures the quality of the ranking of retrieved passages, taking into account their relevance and position in the list.
- BERT-Score: A metric that uses contextual embeddings to compute the semantic similarity between generated and reference texts.
- ROUGE-1: A metric that measures the overlap of unigrams (individual words) between the generated and reference texts. It's particularly useful for assessing the content coverage of the generated responses.

4.2.2 Results: Comparing RAG and Non-RAG with Handwritten Answers

The authors evaluated the generated outputs from the RAG-based and GPT-3.5-based approaches against an expert-curated gold standard. As shown in Table 1, the RAG-based approach outperforms the GPT-3.5-based approach across all three selected metrics (NDCG, BERTScore,Rouge-1). In line with existing research, this highlights the effectiveness of the proposed RAG-based approach in managing domain-specific knowledge.

Table 1: Automatic Comparison: RAG vs. Gold-standard and Non-RAG vs. Gold-standard.

Model	Туре	Nu	NDCG	BScore	Rouge-1
RAG	Simple	10	0.781	0.889	0.345
RAG	Complex	35	0.788	0.887	0.355
RAG	Out-of-S	5	0.750	0.853	0.239
RAG	All	50	0.783	0.884	0.341
Non-RAG	Simple	10	0.763	0.886	0.327
Non-RAG	Complex	35	0.777	0.885	0.340
Non-RAG	Out-of-S	5	0.739	0.853	0.233
Non-RAG	All	50	0.771	0.882	0.327

Additionally, the authors conducted a comparison based on the type of question (complex, simple, or impossible). As illustrated in the same Table 1, in all three categories, the RAG-based approach outperformed the non-RAG-based approach. However, both the aggregate analysis and more detailed comparisons revealed that the improvement of the RAG-based approach over the non-RAG-based approach was only marginal. This suggests that while these metrics provide a general indication of performance, human inspection and evaluation of the generated outputs are also crucial, as discussed in the next section.

4.3 Manual Evaluation

4.3.1 Study Design

To evaluate the models' outputs from anthropological and pedagogical perspectives, this study implemented a human perception experiment with two evaluators and a senior anthropologist who provided expert guidance throughout the evaluation process.

Participants. The two evaluators were from the research team, had computer science backgrounds and university-level experience grading Bachelor's and Master's examinations. To ensure proper assessment of generated answers, both evaluators read 'Argonauts of the Western Pacific' before beginning the evaluation. The senior anthropologist, who holds a PhD in social anthropology and has experience grading anthropology examinations at both Bachelor's and Master's levels, had also read Malinowski's work prior to the evaluation. All participants volunteered their time for this study and received no monetary compensation for their time.

Tasks. The evaluation process presented evaluators with randomly selected questions and their corresponding answers from the gold-standard benchmark dataset. The evaluators assessed the generated answers using the ATGE metric (detailed below) without knowledge of which model produced each answer, minimizing potential bias. In total, participants evaluated 100 questions: 50 generated using RAG and 50 without RAG.

ATGE Metric. To evaluate the generative outputs of the RAG and NON-RAG approach, the authors designed their own assessment metric called ATGE (Anthropological Text Game Evaluation) metric. The latter evaluates language model performance through three components:

- Answer Quality (AQ): Assesses response quality compared to the gold-standard. Scale: 0 (incomprehensible/incorrect) to 5 (perfect).
- Ethnographic Depth (ED): Measures cultural detail richness. Scale: 0 (no cultural context) to 5 (exceptional cultural detail accuracy).
- Honesty Index (HI): Evaluates information accuracy and consistency. Scale: 0 (completely fabricated) to 5 (no hallucinations) with intermediate scores (4=one minor inconsistency, 3=minor inconsistencies, 2 major inconsistency, 1 significant fabrications).

Each component uses a 0-5 Likert scale, with 5 representing the highest score. The final ATGE score

is calculated as the arithmetic mean: ATGE Score = (AQ + ED + HI) / 3 This produces a comprehensive score between 0 and 5, where 5 indicates optimal performance across all dimensions.

4.3.2 Results

The manual evaluation results using the ATGE metric revealed substantial performance disparities between the RAG-based approach and Non-RAG (gpt3.5turbo), as shown in Table 2. The differences manifested across all three evaluated dimensions: overall answer quality, ethnographic depth, and the honesty index.

Table 2: Manual Evaluation using the ATGE Metric.

Model	Туре	Nu	AQ	ED	Н	ATGE
RAG	Simple	10	3.65	3.5	4.5	3.88
RAG	Complex	35	4.19	4.15	4.83	4.39
RAG	Out-of-S	5	0	0	0	0
RAG	All	50	3.66	3.60	4.28	3.85
Non-RAG	Simple	10	2.9	2.9	4.8	3.53
Non-RAG	Complex	35	3.24	3.1	4.43	3.59
Nom-RAG	Out-of-S	5	0	0	0	0
Non-RAG	All	50	2.84	2.75	4.06	3.22

Manual evaluation revealed that the RAG-based system substantially outperformed GPT-3.5 Turbo (the Non-RAG system), with an overall ATGE score of 3.85 versus 3.22. This advantage was most pronounced in two critical dimensions: ethnographic detail (ED score: 3.60 vs. 2.75) and answer qual-ity (AQ score: 3.66 vs. 2.84). As demonstrated in Table 4 (See Appendix) for an example question, the RAG system produced richer responses by incorporating crucial biographic elements, such as Malinowski's theoretical background and field experience, while also providing essential contextual information like the scientific classification of his research. Though the RAG system also showed overall higher honesty scores, this difference was modest (4.28 vs. 4.06), presumably because 'Argonauts of the Western Pacific' had significant representation in GPT-3.5's training dataset.

Detailed analysis revealed that RAG's advantage was most pronounced when handling complex questions (4.39 vs 3.59). This superiority stems from RAG's ability to synthesize detailed information scattered throughout the source material, including chronological events, locations, and character names. While RAG also performed better on simpler questions, the gap was narrower, possibly due to the widespread availability of basic information about Malinowski's work online. Notably, both systems struggled with the five "trick questions" classified as 'Out-of-Scope' in the dataset, suggesting limitations in their deeper contextual understanding.

The results demonstrate RAG's potential for developing culturally nuanced games, particularly in creating what the authors coined as "thick" anthropological games. This concept, inspired by Clifford Geertz's "thick description"(Geertz, 2008) suggests the possibility of games that transcend surface-level cultural representation to offer rich ethnographic content while meaningfully engaging with anthropological theory. RAG's superior performance in handling complex queries and providing detailed ethnographic information makes it particularly suited for such an application.

5 CONCLUSION

In this study, the authors developed a RAG-based prototype designed to facilitate playful exploration of anthropological texts. To evaluate its effectiveness, this study created a handcrafted benchmark dataset comprising 50 questions and answers about Bronislaw Malinowski's anthropological classic 'Argonauts of the Western Pacific'. These questions were categorized into three types (difficult, simple, out-of-scope) to assess the system's performance across varying levels of complexity. Their evaluation methodology combined quantitative analysis of the RAG-based system's outputs against the human-annotated benchmark dataset, while also comparing its performance with a Non-RAG model implementation.

The results demonstrated RAG's superior performance over Non-RAG across multiple metrics. Automatic evaluation showed RAG outperforming GPT on all three metrics (NDCG, BertScore, Rouge-1), while manual evaluation revealed higher ATGE scores (3.85 versus 3.22). The RAG approach produced more accurate responses with stronger ethnographic grounding and fewer hallucinations. These findings validate RAG's effectiveness for text-based LLM games in anthropology and suggest its potential for educational game development.

The implications of this study extend beyond technical achievements to both theoretical and practical domains. Theoretically, the RAG approach enables the creation of what the authors term "thick" anthropological games. Drawing on Clifford Geertz's concept of "thick description" (Geertz, 2008), these games transcend surface-level cultural representations to provide rich ethnographic content while engaging meaningfully with anthropological theory. The results demonstrate that RAG-based systems can effectively support games that approximate this level of ethnographic depth, opening new avenues for anthropological education through interactive and collaborative digital experiences.

But this theoretical contribution may also have significant implications beyond the anthropological domain. One domain that it may affect is the field of intercultural training and communication. The RAG system's ability to maintain ethnographic depth while creating interactive experiences suggests new possibilities for professional cultural education. Organizations could develop dynamic environments where professionals engage in simulated conversations with cultural experts, preserving crucial nuances often lost in conventional training materials.

6 FUTURE WORK

In addition to the improvement idea that has been already laid out, there remain also numerous opportunities for further refinement and optimization of the showcased text-to-text game RAG-driven system. Future iterations could benefit from advances in Retrieval-Augmented Generation technology, such as the Advanced and Modular RAG approaches (Gao et al., 2023), to enhance the quality and depth of the gaming experience. In this context, the findings of Lin et al (Lin et al., 2023) suggest a promising direction by integrating RAG with fine-tuning, which could help identify the optimal combination of these methods to improve the prototype.

Moreover, adopting a multi-modal approach (Yasunaga et al., 2022)—incorporating diverse data types from the original monograph, including information embedded in images—presents additional opportunities to optimize the RAG-based system. Furthermore, this research could expand to evaluate which works across anthropology and other humanities disciplines are most suitable for gamification through RAG-based systems.

ETHICAL CONSIDERATIONS

RAG systems' technical capabilities pose several ethical challenges, particularly in handling anthropological content. Firstly, while efficient at information retrieval, these systems risk oversimplifying cultural nuances and favoring engagement over academic rigor. To address this, the authors of this study added a repository disclaimer emphasizing the tool's role as a supplement to, not replacement for, traditional monograph study.

Secondly, the study's human annotation process, while essential, introduces potential subjectivity. To

enhance assessment validity, annotators consulted a senior anthropology expert when uncertain. Though this expertise helps minimize misinterpretation, inherent evaluation biases may persist.

A third ethical concern relates to privacy and trust relationships in anthropological research. Although published anthropological texts typically anonymize data, some may contain sensitive information shared within specific trust relationships, and the proposed prototype could potentially increase accessibility to such content. To address this concern, the authors deliberately selected a monograph that is over a century old and has been widely circulated and discussed in academic discourse. This choice minimizes the risk of exposing sensitive personal information that might persist within the original text.

AUTHOR CONTRIBUTIONS

Conceptualization: MH, JF; Methodology: MH; Software: MH; Resources, supervision, and project administration: AP; Writing–original draft: MH; Writing–review and editing: MH, SP, AP; Visualization: MH.

REFERENCES

- Ashinoff, B. K. (2014). The potential of video games as a pedagogical tool.
- Aynetdinov, A. and Akbik, A. (2024). Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. arXiv preprint arXiv:2401.17072.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.
- Bouchrika, I., Harrati, N., Wanick, V., and Wills, G. (2021). Exploring the impact of gamification on student engagement and involvement with e-learning systems. *Interactive Learning Environments*, 29(8):1244–1257.
- Browne, C. B. (2008). Automatic generation and evaluation of recombination games. PhD thesis, Queensland University of Technology.
- Cherumanal, S. P., Tian, L., Abushaqra, F. M., de Paula, A. F. M., Ji, K., Hettiachchi, D., Trippas, J. R., Ali, H., Scholer, F., and Spina, D. (2024). Walert: Putting conversational search knowledge into action by building and evaluating a large language model-powered chatbot. arXiv preprint arXiv:2401.07216.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C.,

Gehrmann, S., et al. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

- Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217.
- Freire, M., Serrano-Laguna, Á., Manero Iglesias, B., Martínez-Ortiz, I., Moreno-Ger, P., and Fernández-Manjón, B. (2023). Game learning analytics: Learning analytics for serious games. In *Learning, design,* and technology: An international compendium of theory, research, practice, and policy, pages 3475–3502. Springer.
- Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Togelius, J., and Yannakakis, G. N. (2024). Large language models and games: A survey and roadmap. arXiv preprint arXiv:2402.18659.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.
- Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in entertainment (CIE)*, 1(1):20–20.
- Geertz, C. (2008). Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, pages 41–51. Routledge.
- Hann, C. and James, D. (2024). One Hundred Years of Argonauts: Malinowski, Ethnography and Economic Anthropology, volume 13. Berghahn Books.
- Hoffmann, M. P., Fillies, J., and Paschke, A. (2024). Malinowski in the age of ai: Can large language models create a text game based on an anthropological classic? *arXiv preprint arXiv:2410.20536*.
- Hua, M. and Raley, R. (2020). Playing with unicorns: Ai dungeon and citizen nlp. DHQ: Digital Humanities Quarterly, 14(4).
- Huesmann, L. R. (2010). Nailing the coffin shut on doubts that violent video games stimulate aggression: comment on anderson et al.(2010).
- Järvelin, K. and Kekäläinen, J. (2002). Cumulated gainbased evaluation of ir techniques. ACM Transactions on Information Systems (TOIS), 20(4):422–446.
- Leach, J. W. and Leach, E. (1983). *The Kula: new perspectives on Massim exchange*. CUP Archive.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474.
- Lin, X. V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., et al. (2023). Ra-dit: Retrieval-augmented dual instruction tuning. arXiv preprint arXiv:2310.01352.
- Liu, J. (2023). Building production-ready rag applications.
- Malinowski, B. (2013). Argonauts of the western Pacific: An account of native enterprise and adventure in the archipelagoes of Melanesian New Guinea [1922/1994]. Routledge.

- Manathunga, S. S. and Illangasekara, Y. (2023). Retrieval augmented generation and representative vector summarization for large unstructured textual data in medical education. *arXiv preprint arXiv:2308.00479*.
- Manuaba, I. B. K. (2017). Text-based games as potential media for improving reading behaviour in indonesia. *Procedia computer science*, 116:214–221.
- Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
- Pell, B. (1992). Metagame in symmetric chess-like games.
- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language models as knowledge bases? arXiv preprint arXiv:1909.01066.
- Pia, A. (2019). On digital ethnographies. anthropology, politics and pedagogy (part i).
- Reed, A. (2023). 50 Years of Text Games: From Oregon Trail to AI Dungeon. Changeful Tales.
- Rizki, I. A., Suprapto, N., Saphira, H. V., Alfarizy, Y., Ramadani, R., Saputri, A. D., and Suryani, D. (2024). Cooperative model, digital game, and augmented reality-based learning to enhance students' critical thinking skills and learning motivation. *Journal of Pedagogical Research*, 8(1):339–355.
- Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2023). Ares: An automated evaluation framework for retrieval-augmented generation systems. arXiv preprint arXiv:2311.09476.
- Todd, G., Padula, A., Stephenson, M., Piette, É., Soemers, D. J., and Togelius, J. (2024). Gavel: Generating games via evolution and language models. arXiv preprint arXiv:2407.09388.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
- van Horssen, J., Moreton, Z., and Pelurson, G. (2023). From pixels to pedagogy: using video games for higher education in the humanities. *Journal of Learning Development in Higher Education*, (28).
- Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.
- Wang, A. I. and Tahir, R. (2020). The effect of using kahoot! for learning–a literature review. *Computers & Education*, 149:103818.
- Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. (2022). Retrieval-augmented multimodal language modeling. arXiv preprint arXiv:2211.12561.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

APPENDIX

Table 3: Example Questions from the Q & A Dataset.

Question	Туре
What is the "Kula Ring" and what is its basic function in Trobriand society?	Simple
Describe the political organization of the Kula, including the role of chiefs and their vary-	Difficult
ing levels of power.	
In an alternate history where the Trobriand Islanders developed advanced seafaring tech-	Out-of-Scope
nology and became a dominant Pacific empire, how would their practices have Kula trans-	
formed on a global scale?	

Table 4: Example generated RAG and NON-RAG Answers.

Question	Describe the preface written by Sir James G. Frazer and his assessment of your work and re-
	search methods
RAG	Sir James G. Frazer's assessment of my work and research methods in the preface was highly complimentary. He highlighted that I conducted my research under the best conditions and in a manner calculated to secure the best possible results. Frazer emphasized that I was well-equipped for the task I undertook, citing both my theoretical training and practical experience. He noted my theoretical training demonstrated in my work on the family among the aborigines of Australia and my practical experience based on my account of the natives of Mailu in New Guinea, where I resided for six months. Frazer commended my immersive fieldwork approach in the Trobriand Islands, living among the natives for an extended period, engaging in daily observations, conversations in their language, and direct interactions to gather information first-hand.
	Overall, Frazer's assessment underscored the depth and rigor of my ethnographic methods, em- phasizing the scientific value of the materials I collected and my commitment to understand- ing the social, religious, and economic aspects of the Trobriand Islanders. He highlighted my meticulous preparation and dedication to providing a comprehensive and scientific account of a "savage" people, recognizing the importance of my research in advancing anthropological knowledge.
Non-RAG	Sir James G. Frazer's preface to my book "Argonauts of the Western Pacific" provides a valuable insight into his assessment of my work and research methods. Frazer, a prominent anthropologist himself, acknowledged the significance of my fieldwork in the Trobriand Islands and the inno- vative approach I took in studying the social organization and customs of the indigenous people. He praised my meticulous attention to detail and my focus on understanding the functional as- pects of cultural practices, which laid the foundation for the development of functionalism in anthropology. Frazer recognized the importance of my immersive participant-observation method, where I lived among the Trobriand Islanders for an extended period to gain a deep understanding of their way of life. He appreciated how I emphasized the practical functions of rituals, beliefs, and social structures within the context of the society, rather than viewing them as isolated customs. Frazer's preface highlights his admiration for my groundbreaking research approach, which emphasized the interconnectedness of cultural elements and their role in maintaining social order and stability.