LLMs Can Model Non-WEIRD Populations: Experiments with Synthetic Cultural Agents

Augusto González Bonorino[1], C. Mónica Capra*[2,3] and Emilio Pantoja[4]

[1]Department of Economics, Arizona State University
[2]Department of Economic Sciences, Claremont Graduate University
[3]Center for the Philosophy of Freedom, University of Arizona
[4]Department of Economics and Computer Science, Pitzer College

Abstract

Despite its importance, studying economic behavior across diverse, non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations presents significant challenges. We address this issue by introducing a novel methodology that uses Large Language Models (LLMs) to create synthetic cultural agents (SCAs) representing these populations. We subject these SCAs to classic behavioral experiments, including the dictator and ultimatum games. Our results demonstrate substantial cross-cultural variability in experimental behavior. Notably, for populations with available data, SCAs' behaviors qualitatively resemble those of real human subjects. For unstudied populations, our method can generate novel, testable hypotheses about economic behavior. By integrating AI into experimental economics, this approach offers a proof-of-concept for an effective and ethical method to do exploratory analysis, pilot experiments, and refine protocols for hard-to-reach populations. Our study provides a new tool for cross-cultural economic studies and highlights the potential of LLMs to advance experimental and behavioral research.

Keywords: Behavioral Games, Experiments, Large-Language Models, Small-scale societies, Synthetic Cultural Agents.

JEL Classification: C72, C88, C91, D90

*We thank Matthew Feng for research support. We appreciate feedback and comments from two anonymous referees, Robert Klitgaard, Ernan Haruvy, Daniel Houser, and the late Gary Charness, and participants at the 2025 NBER/CEME Decentralization Conference, the 2024 Barcelona Summer School of Economics, the Economic Science Association, the IAREP seminar series, and the Caltech's Frontier Choice Process Mini-Conference. This research was supported by the Blais Challenge Grant at Claremont Graduate University.
Please send correspondence to Prof. Mónica Capra. Address: 170 E 10th St, Claremont, CA, 91711 USA. E-mail: monica.capra@cgu.edu.

1 Introduction

Large language models (LLMs), such as OpenAI's ChatGPT, have emerged as powerful tools for generating human-like text, but their potential for advancing cultural behavioral research remains largely untapped. This is, in part, because LLMs are pretrained with Western-centric data. In this study, we present a novel methodology that leverages LLMs to create synthetic agents representing non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations. By subjecting these agents to established economic experiments, we provide a proof-of-concept demonstration that LLMs can model non-WEIRD populations. We offer a promising approach for conducting exploratory and preliminary investigations and for generating hypotheses about the behaviors of hard-to-reach populations in an ethical and efficient manner.

The disproportionate representation of WEIRD populations in behavioral research has long been a concern in the social sciences [1]. These populations comprise only about 13% of the global population yet dominate research studies, raising questions about the generalizability of findings to broader human populations. Small-scale societies, in particular, offer valuable insights into human behavior and decision-making, as they more closely represent our evolutionary past [2]. However, studying these populations presents significant challenges, including geographical remoteness, ethical considerations, and methodological inconsistencies across studies [3].

Our approach presents a new methodology to start addressing these challenges by creating Synthetic Cultural Agents (SCAs), LLM-based models that represent specific cultural profiles. Using a combination of web scraping, LLMs, and retrieval-augmented generation (RAG) prompting, we construct cultural profiles for six small-scale societies: the Hadza, the Machiguenga, the Tsimané, the Aché, the Orma, and the Yanomami. We then use these profiles to instantiate LLM agents and subject them to three classic economic experiments: the dictator game [4], the ultimatum game [5], and the endowment effect [6].

Our results reveal substantial cross-cultural variability in experimental behavior, an absence of purely self-interested (*homo economicus*) behavior across all SCAs, and qualitative resemblances to real human populations where data are available. These findings not only generally align with previous research on small-scale societies as described in Henrich, et al. [3] but also extend our understanding to previously unstudied groups.

The alignment of our SCA-generated results with existing human studies validates our methodology as an in-sample proof-of-concept showing the potential of LLMs for culturally relevant behavioral research. This approach, however, does not meet the *data leakage* validation standard set by Ludwig et al. [7], which requires that the validation data not be present in the LLM's training corpus. Nevertheless, we build upon the pioneering work of Horton [8], Hewitt et al. [9], and Mei et al.[10], who demonstrated that LLMs like ChatGPT can produce results qualitatively similar to those of human subjects in experiments and can capture human behavioral norms. Our study, however, advances these previous works by deliberately generating cultural profiles for the target populations of interest and incorporating them into the LLMs for economic experiments. Our methodology allows us to generate cultural profiles, thereby enhancing the simple LLM known to represent WEIRD due to its pretraining on Western-centric data, as shown in the work of Atari et al. [11]. More specifically, we utilize the emerging capability of "in-context learning" [12], which allows us to construct synthetic agents that better reflect the cultural norms, values, and behaviors of specific populations. This method enhances the LLM's ability to represent culturally-specific behaviors in experimental settings.

This study advances the application of artificial intelligence in experimental economics and cultural anthropology, making several key contributions: (1) We introduce a novel methodology for creating synthetic cultural agents (SCAs) using large language models (LLMs), providing researchers with a promising new tool for exploratory research and for piloting experimental protocols, particularly in hard-to-reach populations.[1] (2) We demonstrate that these SCAs can generate responses aligned qualitatively with key behavioral patterns observed in human subjects from the same societies, while also applying this technique to previously unstudied populations. This dual approach is proof-of-concept of the method's potential for preliminary behavioral research, and illustrates how SCAs could be used to anticipate probable behaviors and inform future research design, even where no prior studies exist. (3) We present a multimodal platform for refining experimental protocols, exemplified through an implementation of the endowment effect experiment.

Importantly, this work is not intended to replace human participant research, but rather to complement. Researchers can use SCAs to conduct preliminary investigations, refining their hypotheses and experimental designs before undertaking resource-intensive field studies. While our approach shows promise, we acknowledge its limitations, including potential biases inherent in LLMs and the need for careful validation against human data as described in Ludwig et al. [7]. This study

[1] In a recent working paper, Charness et al. [13] provide a comprehensive exploration of how LLMs can enhance scientific experimentation in the social sciences, highlighting both the opportunities and challenges this integration presents. The authors argue that LLMs can help design experiments, run them smoothly (especially online), and analyze data. The authors also suggest a framework to use LLMs safely while getting the most benefit from them.

contributes to the ongoing evolution of methodologies in behavioral sciences, potentially expanding our ability to study diverse human populations in a more ethical and efficient manner.

2 Methods

The key design principles underlying our methodology are customization and replicability. Specifically, our framework enables researchers to build upon the initial architecture by customizing it to their needs while ensuring that the final model behavior is qualitatively consistent with human behavior reported in the literature. In this paper, we apply the methodology to create cultural profiles of various small-scale societies, instantiate language models instructed to behave as if they were members of the society described by the profile, and subject these synthetic agents to a series of economic games. Through the proposed methods we demonstrate that LLM agents can be used for piloting behavioral studies at lower cost, improving experimental instructions, and studying new and hard-to-reach populations.

The basic framework consists of three key steps, as illustrated in Figure 1. We begin by building a knowledge base. We developed and tested three methodologies: Direct Prompting, Self-Ask with Search [14], and Search + Retrieval Augmented Generation (RAG) [15]. In Step 2, we utilize an LLM to generate a cultural profile based on the relevant context retrieved in Step 1. We tested Anthropic's Claude 3 Opus and GPT-4o to generate said profiles and found Claude's results to be both more coherent and detailed. However, the methodology is model-agnostic and can be applied to any language model, whether closed-source or open-source.[2]

In Step 3, we use the cultural profile to instantiate a new LLM with a system prompt that instructs the model to respond as if it were a representative agent of the tribe sharing similar preferences and viewpoints. The resulting customized LLM agent serves as our synthetic representative agent, which we subsequently subject to a series of economic experiments (see Section 2.3).[3] We use ChatGPT 3.5 because of its lower cost per token and faster inference.[4]

[2]The independence from proprietary models is an upside of our methodology. We plan to extend our analysis by benchmarking open-source models available via HuggingFace to compare their performance with that of closed-source counterparts.

[3]The cultural profile and behaviors generated also depend on hyper-parameters (i.e., temperature) and the inherent stochastic nature of generative models.

[4]ChatGPT3.5 is comparatively less powerful than more advanced models, but the RAG prompting approach lowers the required out-of-the-box capabilities of the LLM by explicitly defining the text (i.e., profile) to reference and leveraging in-context learning. We tested ChatGPT4 and Claude 3 Opus as more powerful alternatives to ChatGPT3.5 in the experimental decision step and observed no significant gains in reasoning or coherence. This is why we opted for the least expensive option.

We create synthetic representative agents representing six distinct small-scale societies: the Hadza, the Orma, the Tsimané, the Machiguenga, the Aché, and the Yanomami (see Section 2.2.1). These societies were chosen for their diversity in economic systems, social structures, and geographical locations, as well as the varying availability of existing experimental data, allowing us to both validate our approach and extend insights to previously unstudied groups.

For each of these six tribes, we construct a comprehensive cultural profile for the synthetic agent. In constructing the cultural profiles, we considered key socio-cultural-economic factors including lifestyle, cultural practices, economic systems, political ideologies, social organization, kinship structures, and core values [16, 17]. These comprehensive profiles inform the behavior of our SCAs, which are then prompted to participate in established economic experiments such as the dictator game, ultimatum game, and tests for the endowment effect.

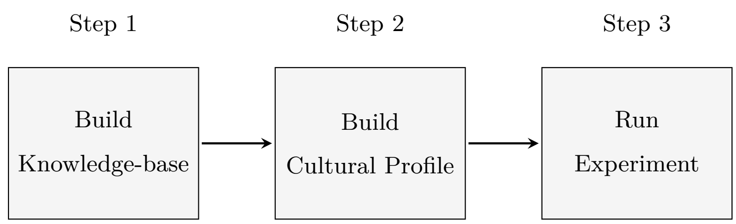

Figure 1: Framework for Constructing Experiments with Synthetic Agents

Note: Step 1. Build Knowledge-base: Directly prompt or fetch publicly available information and upload local documents to store as the LLMs knowledge, which represents the context for building a cultural profile. **Step 2. Build Cultural Profile:** If Step 1 is RAG, retrieve the top k most relevant chunks of documents in the knowledge base, and inject them in the prompt as context to reference for generating the cultural profile. **Step 3. Run experiment:** A new customized LLM is instantiated with the instruction to respond as if it were a member of the society described by the cultural profile generated in Step 2. The custom LLM is then subjected to a series of economic games/experiments through direct prompting.

2.1 Build Knowledge-base and Cultural Profile

We consider three methodologies for the implementation of Steps 1 and 2 represented in Figure 1: Direct prompting, LangChain Agents, and Search + RAG. Each method represents a particular prompting technique, which differs in how the contextual information is fetched and leveraged to construct the cultural profile. The generated cultural profiles using direct prompting can be found downloaded from our GitHub repository.

2.1.1 Direct Prompting

The first methodology, Direct Prompting, consists of simply prompting the LLM to generate a profile by primarily sampling information from its training data, conditioned on a list of relevant factors (input as a list of keywords to consider). Table 1 shows an example of the prompts used to build a cultural profile of the Hadza. A detailed description of this technique is included in Appendix A.1.1 and the code is available through our GitHub repository.

```
# Profile builder
system\_prompt = "You're a helpful assistant that aids
                 in constructing detailed and comprehensive cultural profiles"
tribe\_to\_search = "Hazda"
relevant\_factors = ["lifestyle," "culture", "social organization," "kinship,"
                    "economic system", "political ideologies present", "values"]

# Profile builder task prompt (this outputs the cultural profile)
prompt = system\_prompt \
        + f"Please construct a profile on the {tribe_to_search}. " \
        + f" The profile must cover the following
            socio-economic relevant factors {relevant_factors}.
            Proceed step by step."
```

Table 1: Example of Direct Prompting

Note: Code snippet of Direct Prompting to build a cultural profile. The table shows the system prompt and the prompt for creating a cultural profile for the Hadza.

Direct prompting is the methodology most closely aligned with the approaches used by both Horton [8] and Mei et al. [10]. In this scenario, a cultural profile is built solely based on the training data the LLM has access to and a specific temperature chosen before generation. Although important insights can be gained from LLMs prompted to acquire a cultural profile, there are problems with this approach.

To begin, LLMs are pre-trained language models, which means that the data to build the profile are static at the time of usage. Indeed, LLM knowledge base contains all the information on the internet up to a training cutoff date. Thus, leaving out the possibility of referencing the most current available information when building the cultural profile.

Furthermore, LLM generations can be unreliable due to two primary factors: insufficient reference sources and hallucinations. The latter refers to outputs that, while coherent, are factually incorrect. These issues can result in biased or inaccurate cultural profiles. There is also little control over exactly what part of the embedded knowledge space the model references during generation. Nevertheless, direct prompting is the fastest, least expensive, and easiest approach because it requires no additional steps to curate context.

2.1.2 LangChain SelfAsk with Search Agent

LangChain's "SelfAsk with Search" agent is an LLM model equipped with tools that enable it to automate tasks [18]. The agent is built into the LangChain Python module, thus it is freely and easily accessible. This methodology relies on fetching relevant information autonomously from the web and generating a Cultural Profile (steps 1 and 2) by combining the self-ask prompting method with a Google search tool. The search tool endows an LLM with the capability of browsing the internet given a query [14]. The self-ask prompt builds upon Chain-of-Thought [19] by applying a Reason-Act [20] framework. This guides the model's chain of reasoning via a series of automatically constructed follow-up questions constrained by the relevant socio-economic factors specified in the prompt. Table A.1 shows an excerpt of the Agent's execution chain given the prompt shown in Table 1. Once no more follow-up questions are generated, or a maximum number of iterations is reached, the LLM parses through the intermediate answers and generates a cultural profile. Our GitHub repository includes code along with detailed explanations of our methodology.

The LangChain's agent allows us to update the knowledge base of the LLM through search capabilities of updated information, and guide the LLM to retrieve relevant information from its searches. Thus, this method reduces hallucinations [21] and updates information processed by the LLM beyond the cutoff date. However, the method has limitations. First, the sources referenced were not evident and no immediate way of constraining the sources to visit existed at the time of writing this paper.[5] Second, the profile was often too short and information was tightly compressed (see Appendix A.1.2). Third, runtime scales badly with number of follow up questions. Fourth, the number of tokens used is significantly higher. Therefore, this architecture provides a more factually consistent generation at the cost of runtime and tokens.

2.1.3 Search + RAG Methodology

Enhancing the system prompt with information from the web is beneficial, but the lack of transparency makes using built-in agents a "black box" effect. This opacity makes it difficult for researchers to dissect the reasoning process or fully comprehend how these agents arrive at their outputs. Retrieval-augmented-generation (RAG) [15] is a technique that combines information retrieval techniques to augment the prompt of an LLM with relevant context to improve the quality and factual accuracy of generated text. In our third methodology, Search + RAG, we decouple the

[5] Major upgrades have been implemented to LangChain and direct model providers' APIs, so this limitation bears little weight in late 2025.

```
Follow up: What is the lifestyle of the Yanomami?

Intermediate answer: The Yanomami are the largest relatively isolated tribe in South America.
    They live in the rainforest and mountains in northern Brazil and southern ... Mar 2, 2024
    ... Yanomami, South American Indians, speakers of a Xiriana language, who live in the
    remote forest of the Orinoco River basin in southern ... Aug 9, 2022 ... Today, the
    Yanomami - who number about 29,000 - say they are at serious risk of losing their lands,
    culture and traditional way of life. The ... Nov 15, 2018 ... The Yanomami diet, low in
    fat and salt and high in fiber, consists of such items as plantains, cassavas (a root
    vegetable), fruit, and meat- [...]

Follow up: What is the culture of the Yanomami?

Intermediate answer: The Yanomami are one of the most numerous, and best-know, forest-dwelling
    tribes in South America. Their home is in the Amazon rainforest, among the hills...Mar 2,
    2024... Yanomami, South American Indians, speakers of a Xiriana language, who live in the
    remote forest on the Orinoco River basin in southern...Aug 9, 2022 ... Today, the Yanomami
     - who number about 29,000 - say they are at serious risk of losing their lands, culture
    and traditional way of life. The ... The Yanomami, also spelled Yanomamo or Yanomama, are
    a group of approximately 35,000 indigenous people who live in some 200-250 villages in the
     Amazon [...]

Follow up: What is the economic system of the Yanomami?

Intermediate answer: [...]
```

Table 2: Example of a LangChain Agent Execution Chain

Note: Code snippet depicting the LangChain agent's chain of reasoning. The agent parses the {relevant_factors} such as lifestyle, and culture one at a time. In the model, once no more follow-up questions are generated, or a maximum number of iterations is reached, the LLM parses through the intermediate answers (e.g., "The Yanomami are the largest...") to generate a cultural profile.

search and retrieve tasks from the LLM providing more control and transparency about the source of information for building the knowledge base and cultural profile. Figure A.1 in the Appendix illustrates this third methodology.[6]

For building the knowledge base of a tribe, we start with the search query *"What characterizes the (name_of_tribe) tribe?"* The search task runs a Google search and returns the URL links of the top k results for the query. The links are fed to a function that scrapes and parses the textual information in each source to store as context. This step is done asynchronously to improve runtime performance. The retrieval task is achieved by splitting the scraped documents into chunks of a

[6]Ideally, we would like to insert as much context as possible to increase the likelihood of generating a high-quality cultural profile. However, there are two drawbacks to this approach. On one hand, we are constrained by the context window of the particular LLM being deployed. On the other hand, it increases the risk of falling for the "Lost in the Middle" trap first documented by researchers at Anthropic [22]. They note that this can be corrected with modifications to the original prompt, but simple prompting has problems that may be difficult to discover by looking at the qualitative data. Hence, we need a way of retrieving only the most relevant parts of the scraped context to control for the length of the prompt and prevent missing relevant information in the body of the documents.

fixed length, indexing them in a vector store, and instantiating a retrieval model that ranks chunks based on a similarity measure (e.g., dot product). We implement each step (search, fetch, retrieve, and rank) manually, thus allowing for full customization of the RAG architecture. The retrieved indexed chunks are injected into the profile-building prompt as context for building the cultural profile.

```
results = tool\_sources.run(prompt)
results
    [{'title': 'Hadza',
        'link': 'https://www.nationalgeographic.org/encyclopedia/hadza/',
        'snippet': 'Oct 19, 2023 ... The Hadza are a modern hunter-gatherer people living in
            northern Tanzania. They are considered one of the last hunter-gatherer tribes in\
            xa0...'},
     {'title': "This East African Tribe's Lifestyle Is Depression-Proof | by Will ...",
       'link': 'https://medium.com/illumination/living-to-survive-are-societal-complexities-
           destroying-our-lives-430aa9620a6f',
       'snippet': 'Jun 20, 2023 ... What we can learn from David Choe \& the Hadza tribe. ...
           Have you ever felt as if life is just too complicated? Not a big Joe Rogan guy, but
           \xa0...'},
    {'title': 'Hadza people - Wikipedia',
         'link': 'https://en.wikipedia.org/wiki/Hadza_people',
         'snippet': 'The Hadza, or Hadzabe (Wahadzabe, in Swahili), are a protected hunter-
             gatherer Tanzanian indigenous ethnic group from Baray ward in southwest Karatu\
             xa0...'},
    {'title': 'Is The Secret To A Healthier Microbiome Hidden In The Hadza Diet?',
        'link': 'https://www.npr.org/sections/goatsandsoda/2017/08/24/545631521/is-the-secret-
            to-a-healthier-microbiome-hidden-in-the-hadza-diet',
        'snippet': "Aug 24, 2017 ... Some species of bacteria in our intestines are
            disappearing. Can we reverse the microbial die-off? The food eaten by Tanzania's
            Hadza tribe\xa0..."},
     {'title': 'Helping the Hadza Protect Their Homeland',
        'link': 'https://www.nature.org/en-us/about-us/where-we-work/africa/stories-in-africa/
            the-hadza-helping-hunter-gatherers-protect-their-homeland/',
         'snippet': 'Northern Tanzania is home to the Hadzabe, one of the last remaining hunter
             -gatherer tribes on Earth. Known for shunning material possessions and social\xa0
             ...'}]
```

Table 3: Example of Sources Retrieved

Note: Code snippet depicting the sources utilized for the cultural profile of the Hadza. Unlike Direct Prompting and LangChain Agent, with the Search + RAG methodology, we can see the sources or documentation that generate context for the profile and the instantiation of the synthetic agent.

We can control the chunking method, embedding model, choice of vector store, and retriever's parameters. For our experiment we use 2000 words chunk size with 200 words overlap, the BGE open-source embeddings model [23], the open-source Chroma DB vector store to manage our text embeddings and perform similarity searches, with $k = 10$. The RAG chains these two components (i.e., Search and RAG) together to enhance the original prompt used to generate a cultural profile with domain-specific and relevant context. That is, the LLM is *instructed to pay attention* to the retrieved information and combines it with the user prompt to generate an answer. A more detailed explanation of the methodology can be found in Appendix A.1.3 and our GitHub repository.

This methodology reduces hallucinations because the augmented data is stored with relevant and up-to-date context data. It also enables us to include online and offline information, and the modularity of the architecture allows for easy customization. For example, one can fit proprietary data, or constrain the information stored. The methodology also enhances transparency by allowing us to track sources. Table A.2 shows a snippet of the online sources that the model uses to build a profile of the Hadza.

We find that the generated profiles are much more consistent and of overall higher quality, and it takes less time to run this model than the previous architecture. That said, Search + RAG requires more programming effort. Table A.3 summarizes the advantages and disadvantages of each methodology, offering a clear and direct comparison across various critical dimensions such as runtime, reliability, information freshness, transparency, and customizability. Detailed cultural profile outputs for each of the six tribes, reflective of the distinctions captured by these methodologies are available in our GitHub repository.

The combination of reduced hallucinations, currency of information, transparency, and customizability makes Search + RAG a robust and flexible tool for creating accurate and comprehensive cultural profiles, which serve as the foundation for our Synthetic Cultural Agents (SCAs).

2.2 Creating Synthetic Cultural Agents and Running Experiments

Once we have generated comprehensive cultural profiles using the Search + RAG methodology, we use these profiles to instantiate Synthetic Cultural Agents (SCAs) capable of participating in economic experiments. This process involves two key steps: (1) instantiating the SCA with the cultural profile, and (2) subjecting the SCA to experimental tasks.

2.2.1 Small-scale Societies

We created profiles of six small-scale societies: the Hadza, the Machiguenga, the Tsimané, the Aché, the Orma, and the Yanomami. These societies were chosen for their diverse economic and social organizations, as well as the varying availability of existing experimental data. Table 4 provides an overview of the economic and social organization of these tribes and notes the presence or absence of experimental studies involving any of our three games. This diverse selection allows us to both validate our approach against existing data and extend insights to previously unstudied groups, such as the Yanomami.

SS-Society	Description	Dict. Game	Ult. Game	End. Eff.
Hadza	Hunter-gatherers from northern Tanzania with strong sharing ethics.	✓[24]	✓[25]	✓[26]
Machiguenga	Subsistence horticulturalists from the Peruvian Amazon showing cooperation primarily at the family level.	n.d.	✓[25, 27]	n.d.
Tsimané	Forager-horticulturalists from the Bolivian Amazon with strong sense of economic independence at the level of the nuclear family and extended household.	✓[28]	✓[25]	n.d.
Aché	Hunter-gatherers from Paraguay with strong norms of sharing catch equally and cooperating among households.[29]	n.d.	✓[25]	n.d.
Orma	Pastoral-nomadic community from Kenya that emphasizes communal sharing and cooperation.	✓[30]	✓[25, 30]	n.d.
Yanomami	Forager-horticulturalists from the Venezuelan and Brazilian Amazon that value cooperation and equitable distribution of resources.[31]	n.d.	n.d.	n.d.

Table 4: Economic Organizations, Sharing Norms and Experimental Studies of Six Tribes

Note: The table shows the economic organization and sharing norms of the six small-scale societies for which we generate a cultural profile for our synthetic agent. The third - fifth columns show whether there is experimental data for the tribe; ✓= available data; n.d. = no documented studies.

2.2.2 Instantiating Synthetic Cultural Agents

To create an SCA, we use the cultural profile as a system prompt for a large language model (in this case, ChatGPT 3.5). The system prompt instructs the model to respond as if it were a representative member of the specific tribe, sharing similar preferences, viewpoints, and decision-making processes. This approach leverages the model's ability to condition its outputs on the provided context, effectively creating a digital agent that embodies the cultural characteristics of the target population. Details on prompt structure are available in the Appendix A.2 and our GitHub repository.

We chose ChatGPT 3.5 for this task due to its balance of performance and efficiency. Recent studies have shown that Retrieval-Augmented Generation (RAG) can significantly reduce the required size of the language model by leveraging in-context learning [12], allowing for the use of less powerful models without compromising performance.

2.3 Run Experiment

Our synthetic agents were subjected to three experimental tasks, all of which have been extensively studied in laboratory conditions and the field:

1. The Dictator Game - A player (the dictator) splits an endowment with another player [5].
2. The Ultimatum Game - A player proposes to another player a split of an endowment. If the proposal is rejected, both players earn nothing [4].
3. The Endowment Effect - The tendency to value items more once they are possessed [6, 32].

To run the experiments, we set up two parameters: an experiment prompt and a task prompt. The experiment prompt specifies the decision environment for the synthetic agent, while the task prompt specifies the agent's choice set. We base these prompts on their respective standard approaches in existing literature (see Appendix A.2).

In the dictator and the ultimatum games, we use endowments equivalent to a day's wage to simulate the financial conditions participants encountered in the field settings as described in [3]. We employed the strategy method, asking the SCAs to respond with "yes" or "no" to a series of contingent splits ranging from 0% to 100% of the endowment in 10 percentage-point increments. Table 5 shows the experimental and task prompts for the ultimatum game. The prompts for all experiments can be found in Appendix A.2.

To capture within-subject variation and increase the robustness of our results, we repeated each experimental task 100 times for each SCA. This approach allows us to leverage the language model's inherent stochasticity while ensuring consistent behavior across trials. As Mei et al. [10] and Horton [8], we do not model population variance. The SCA is a representative agent of the population we model.

In addition to the experimental instructions, we prompted the synthetic agent to provide a step-by-step explanation of the rationale for their "yes/no" choice. This additional data allows us to verify the agent's understanding of the task and gain insights into the potential cultural factors (reflected by choice of words generated conditional on the given profile) influencing their decisions. We believe this provides evidence that the probability distribution over tokens is being influenced by the profile and prompt in such a way that it nudges the model to choose words more likely to be relevant in the given cultural context. This does not imply the cultural construct is a perfect representation, nor that this is the best way to steer the models to represent a desired persona, rather we argue that inspecting the explanations are useful for checking if the scales of payoffs, structure of the game, and task were properly "understood" by the SCA. This is akin to Ross et al [33] “competence tests.” By prompting the reasoning process, we are able to verify that ChatGPT3.5 can understand endowment splits and perform budget calculations accurately. Responses for each SCA and for each game can be found in our GitHub repository.

We use a temperature parameter to control the randomness of the model’s text generation. Lower temperatures make responses more deterministic (close to greedy sampling), while higher temperatures increase variability by giving more weight to less likely words. Following Mei et al. [10], we fix the temperature at 1, the default setting for ChatGPT’s API. This setting provides a balance between response randomness and coherence in the model’s outputs by modifying the probability distribution of the next token sample.[7]

Finally, for the endowment effect, we took a different approach. We build upon the methodology discussed in Section 2.1.3 to create a user-friendly multimodal platform for testing the tendency to value things more once they are possessed. As implemented in the field with Hadza tribesmen, the endowment effect had a visual component. In one of their experiments, Apicella et al. [26] placed endowed/exchange items on the ground in front of the Hadza for them to see and choose. In line with this, when developing the endowment effect experiment with SCAs, we employed a two-step multimodal design. First, we generated two items as objects via text-based chat. Next, we tested for the endowment effect by displaying images of the two items to the chatbot, thereby integrating both textual and visual elements into the experimental framework. See Appendix A.2.3. Our code and documentation can be found through our GitHub repository.

3 Results

3.1 Dictator Game

As far as we know, the Dictator Game (DG) was conducted in the field with the Orma, the Hadza, and the Tsimané. We have not found any references to published DG experiments with members of the Aché, Machiguenga, or Yanomami tribes. Among the societies with available data,

[7] It is possible to run our experiments with varying temperatures based on set increments of 0.1. In fact, for robustness, we tested this using several ranges and averaged out the responses across temperatures. We report results for temperature = 1 as it is the default for API usage of ChatGTP.

```
system\_prompt = f"""You are a member of a tribe with the following characteristics. You must
    answer as if you were a representative member of such tribe.
       -----
       {profile}
       -----
You and other members of the same tribe have been selected to participate in an experiment.
   """

if proposer:
   return system\_prompt + "\n\n\n" + (f"You have been chosen to be a proposer of an offer to
       distribute \${amount_endowed},"
       f"which is equal to a day's worth of work, between yourself and another {'tribe member'
            if len(profile) > 10 else 'player'} whose identity is not disclosed. "
       f"If the other {'tribe member' if len(profile) > 10 else 'player'} accepts the offer,
           you keep the difference between \${amount_endowed} and your offer. "
       f"If the other {'tribe member' if len(profile) > 10 else 'player'} rejects the offer,
           you and the other {'tribe member' if len(profile) > 10 else 'player'} get nothing.
           "
       f"Offering exactly \$5 represents an equal split."
       f"\n\nGiven this, would you offer \${offer_percentage} out of \${amount_endowed}?"
       f"You must answer 'Yes' or 'No'.."
       "Start with your answer, followed by a separator token [EXP], and then briefly justify
           your reasoning step by step.")
else:
   return system\_prompt + "\n\n\n" + (f"You have been chosen to be a recipient of an offer
       from another {'tribe member' if len(profile) > 10 else 'player'} whose identity is not
       disclosed. "
       f"The proposer was chosen to distribute \${amount_endowed} between the proposer and you
           ."
       f"You can either accept or reject this offer. "
       f"If you accept the offer, you get what is proposed and the proposer gets the
           difference between \${amount_endowed} and the offer."
       f"If you reject the offer, you and the other get nothing."
       f"An offer of exactly \$5 represents an equal split."
       f"\n\nYou are offered \${offer_percentage} out of \${amount_endowed}. "
       f"Do you accept this offer? You must answer 'Yes' or 'No'."
       "Start with your answer, followed by a separator token [EXP], and then briefly justify
           your reasoning step by step.")
```

Table 5: Ultimatum Game Experiment and Task Prompt

Note: Code snippet depicting the ultimatum game experimental prompt and the task prompt for the proposer and the recipient of an offer. We use the strategy form as the proposer is asked to accept or reject a set of offers and the recipient is asked to accept or reject response contingencies. We also request the agent to provide a brief step-by-step justification for its reasoning.

the Tsimane' of the Bolivian Amazon demonstrates a relatively generous sharing norm (modal offer of 50%), although there is significant variation based on village membership [28]. The Orma of Kenya show similarly high offers, influenced by factors such as market integration and community size [30]. In contrast, the Hadza of Tanzania exhibits lower offers, with a surprisingly low modal offer of 0-10%. Interestingly, this occurs despite their strong sharing ethic, with lower offers particularly observed in smaller camps [24]. For the Machiguenga, Aché, and Yanomami societies, no specific data on DG behavior are available, highlighting the need for further research in these populations. Table C.1 summarizes the observed dictator offers for the three previously studied tribes, along with researchers' key observations on the decisions made in the DG by each tribe.

In our DG experiments, a synthetic agent representing each tribe was prompted to either accept or reject a proposed split ranging from 0% to 100% in 10% increments. We repeated this process 100 times for each agent, allowing us to capture within-subject variation by taking advantage of the inherent stochasticity of the LLM, and increasing the likelihood of consistent answers across trials.[8]

Figure 2 shows the entire distribution of dictators' acceptances and Table C.3 shows the number of accepted offers across different offer rates (0% to 100% in 10% increments). The results show that the decision to accept or reject an offer rate is dependent on the tribe affiliation, suggesting that different cultural or social factors might influence the acceptance rates observed in each tribe (CMH $M^2 = 27.480, p - value < 0.001$).[9] Indeed, the synthetic agents modeled after horticultural societies that primarily cooperate at the family level, like the Machiguenga or the Tsimané, exhibit a different distribution of offers than those modeled after hunter-gatherers like the Aché.

Figure 3 depicts the acceptance rates of zero offers proposed to the synthetic dictator representing different tribes, including ChatGPT, which serves as the WEIRD benchmark. The figure shows variability across cultures in 0% offers, with the horticulturalists who have a sense of economic independence at the family level displaying more self-interested behavior.[10] For instance, the synthetic Machiguenga agent responded "yes" 20% of the time to the question: *would you offer \$0 out of the amount endowed to the other player?* In contrast, the Aché agent accepted the \$0 offer only 3 out of 100 times (p < 0.001).[11]

[8] The "yes/no" approach mimics the strategy method used in human-subject experiments, whereby the decision-maker is asked to make a binding choice for each contingency of the other's action before observing that action (see Selten [34]. The strategy method has also been implemented when studying small-scale-societies the subject pool is limited and the strategy method lets you get a complete response profile for every participant [35].

[9] We applied the Cochran-Mantel-Henszel (CMH) to determine if there is a significant association between group membership and acceptance decisions across different offer levels. Given the structure of our data (i.e., six independent groups, each asked to provide 100 binary responses (accept/reject) for each offer ranging from 0% to 100% in 10% increments), we used the CMH test to control for the varying offer levels while assessing the overall association between tribal membership and decision. By treating each offer level as a separate stratum, the CMH test provides a nonparametric method that accounts for the multiple offer levels. The CMH test evaluates whether the odds ratio of acceptance versus rejection is the same for all groups (tribes) across all strata (offer levels).

[10] We conducted the Chi-Square Test for Independence to examine the association between the different synthetic tribes and their decisions to accept or reject an offer rate of 0%. The results indicate that the proportion of 0% offers differs significantly across the six tribes ($\chi^2(5) = 38.255, p < 0.0001$). See Table C.4 in the Appendix.

[11] To examine the differences in acceptance rates between each pair of groups, we performed Fisher's Exact Test and adjusted the p-values for multiple comparisons. See Table C.5 in the Appendix.

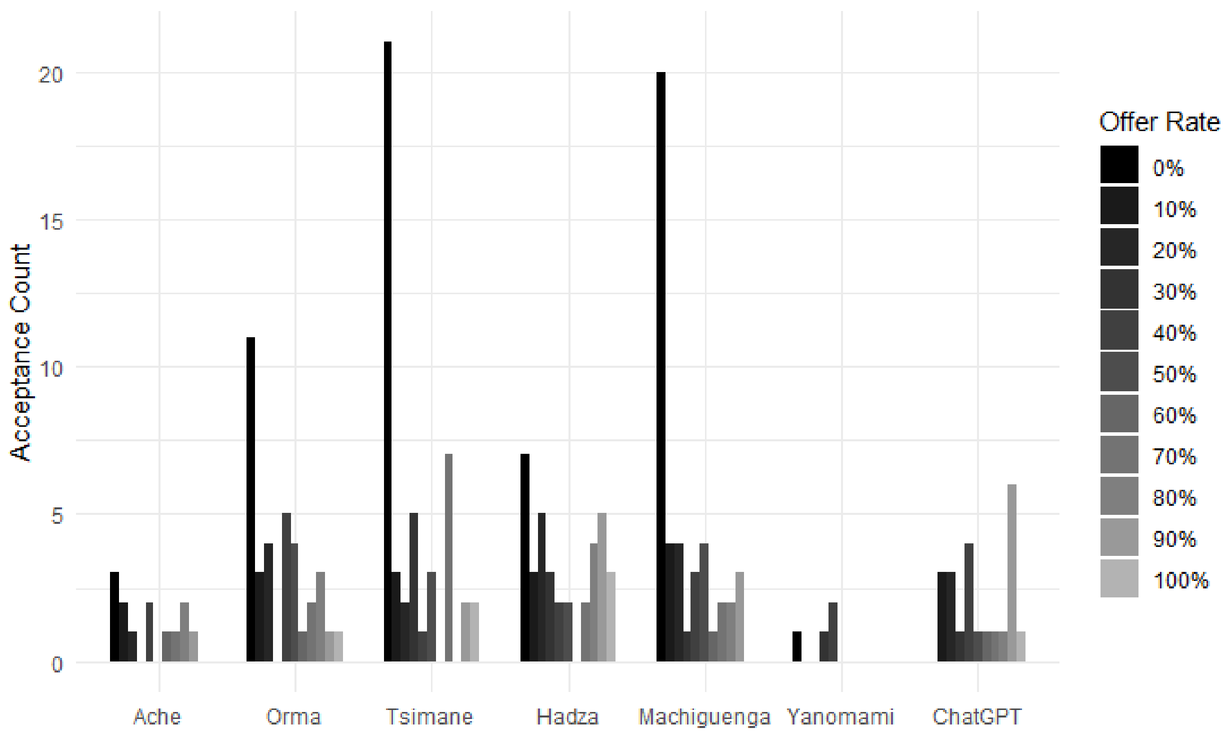

Figure 2: Dictator's Accept Count to Range of Offers by Tribe and ChatGPT

Note: The figure shows the number of "yes" responses by synthetic tribesmen to proposed offers ranging from 0% to 100% in the Dictator Game. Each task (i.e., responding "yes" or "no") is repeated 100 times. ChatGPT represents the WEIRD benchmark. The figure shows large variability across cultures.

Interestingly, ChatGPT WEIRD benchmark rejected all zero offers across all iterations. While this finding contradicts the prevalence of zero modal offers observed in student populations [36], it aligns well with Mei et al.'s findings [10] that ChatGPT tends to avoid selfish behavior and exhibits greater generosity compared to human subjects in experimental settings.

The Yanomami tribe's offering pattern differs significantly from all other tribes except the Aché. Although both exhibit low acceptance rates for 0% offers, the Yanomami overall offer rates are low and compared to other tribes, with the Yanomami accepting offers only 4 times out of 1,100 trials.[12]

The above results suggest that the cultural profiles we generated with our methods play a significant role in influencing dictator game behavior. This finding is consistent with previous research with human subjects, which show that social and economic arrangements help explain variations in dictator game behavior across different populations. For example, studies by Gurven [28], Carpenter [37], and Engel [38] have found that factors such as market integration, community size, and local norms of fairness can affect how individuals allocate resources in the dictator game.

3.2 Ultimatum Game

Previous research on economic behavior in small-scale societies has revealed important variations in ultimatum game (UG) outcomes. A study by Henrich et al. [25] found that both the Machiguenga of Peru and Orma in Kenya made low offers with rare rejections. In contrast, the modal Aché offers were high and there were no rejections. The Tsimané showed no rejections as well, even for low offers, contrasting sharply with the Hadza, who exhibited high rejection rates (43% modal for offers $\leq$ 20%)[2] (see Table C.2).

The differences of behavior in the UG are attributed to factors such as market integration, community size, and cultural norms of sharing and fairness [37]. The Machiguenga and Tsimané, for example, cooperate at the family level and it appears that the anonymity of the players in the UG removes fairness considerations. The Aché regularly share meat, which they distribute equally among all the households, irrespective of which hunter made the catch. In contrast, the Hadza's high rejection rates may be influenced by their tightly-knit communities, where fairness and equity norms are strongly enforced.[13]

[12] We can only speculate on why the Yanomami SCA would behave this way. In the Yanomami, exchanges create ongoing relationships and obligations; the arbitrary division of resources between people with no relationship could be seen as morally wrong. It is possible that playing the games establishes relationships that they may not want; thus, they reject everything.

[13] In small-scale societies, low offers and high rejections are linked to a concept known as "tolerated theft" [39]. In these societies where resources are limited and sharing is expected, individuals may tend to make low offers to hold onto their resources, while also feeling entitled to a fair share of others' resources, leading to the rejection of low

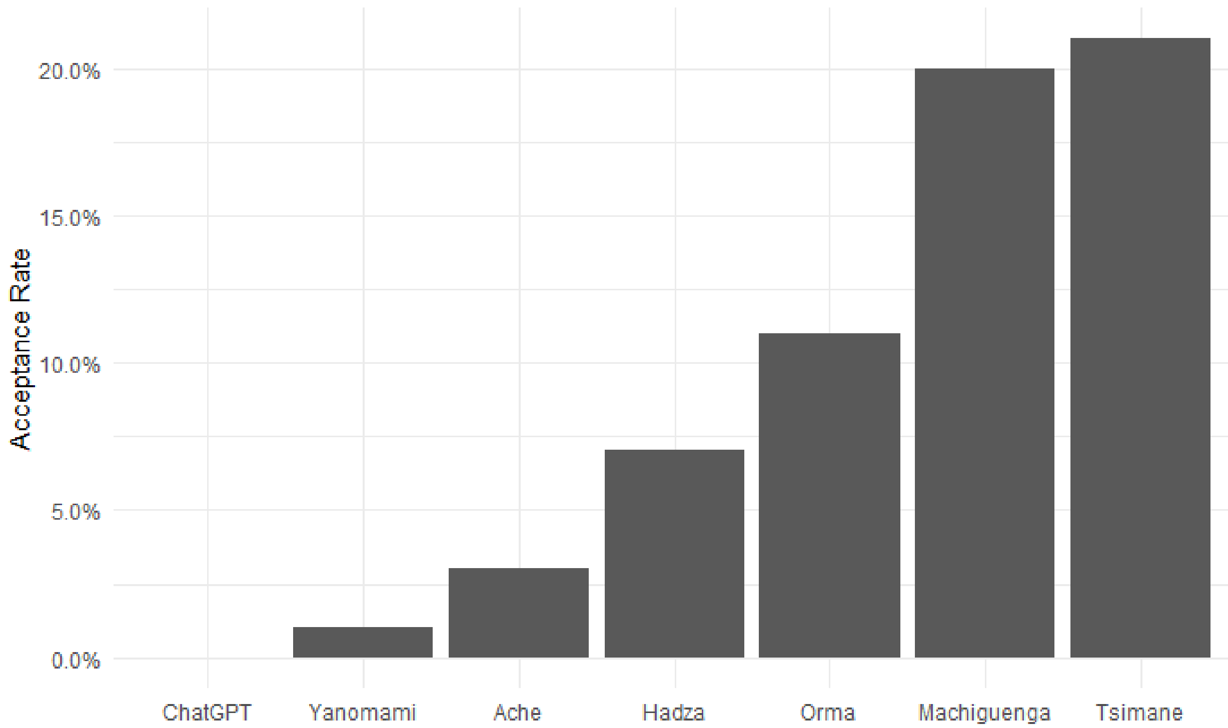

Figure 3: Percentage of Zero Offers by Tribe

Note: The figure shows the percentage of "yes" responses by synthetic tribesmen to offering 0% of the endowment to another player in the Dictator Game. Each task is repeated 100 times to create within-subject variation. ChatGPT said "no" to all contingent zero offers across all 100 iterations.

Figure 4 shows the number of "yes" responses by synthetic tribesmen in the role of proposers to offers ranging from 0% to 100% in the UG. The acceptance counts out of 100 iterations for each offering level are shown in Table C.6. The data demonstrate that among the synthetic tribesmen, the Yanomami exhibit behavior most closely resembling *homo economicus*; all other synthetic tribesmen, including the Hadza, made generous offers. For most synthetic tribes and ChatGPT, the modal offers were 60% of the endowment, which is higher than the modal offers observed in experiments with human participants, as described in Table C.2. Despite of this, we found a significant association between group membership and the proposer's decision, accounting for the different offer levels (CMH $M^2 = 60.796, p < 0.001$).

Figure 5 shows the recipients' decision to reject (i.e., "no" responses) offers ranging from 0% to 100% of the endowment. In contrast to observations with human tribesmen, synthetic tribesmen tend to have high rejections. However, the rejection rates show variability across tribes, largely following observed patterns of behavior in human subjects qualitatively.

offers.

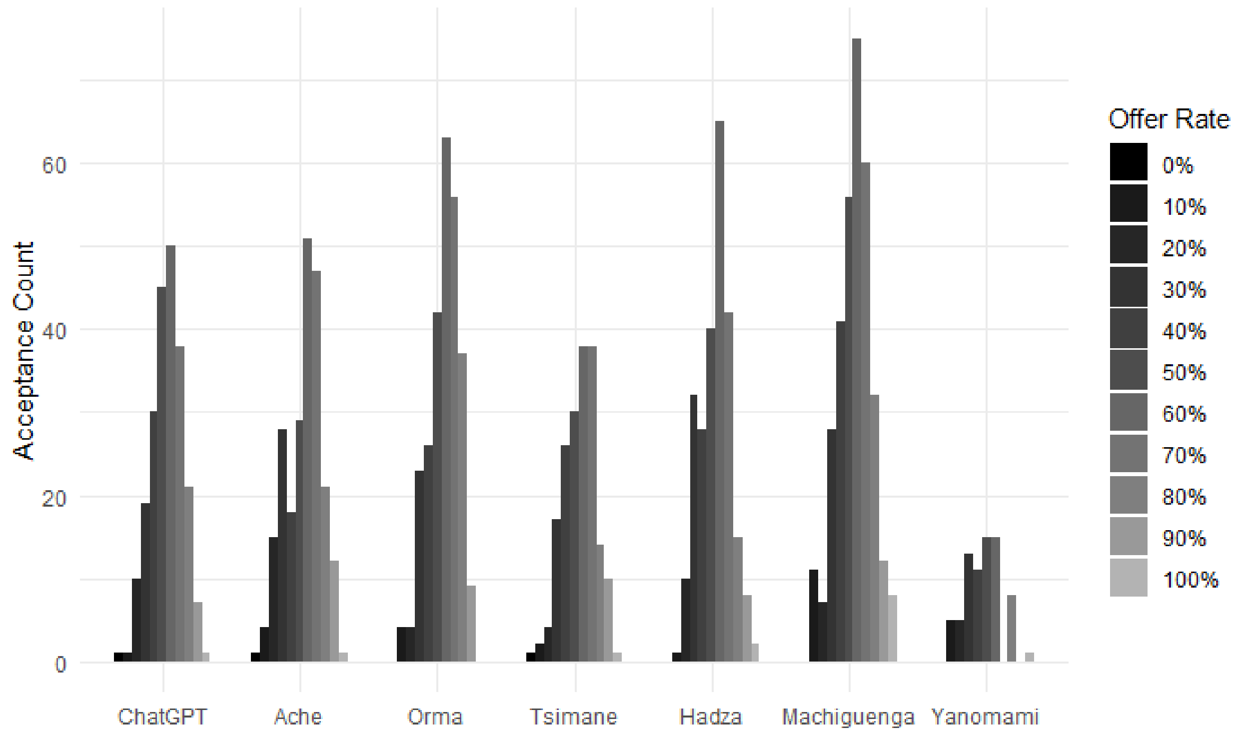

Figure 4: Proposer Accept Count to Contingent Offers by Tribe and ChatGPT

Note: The figure shows the number of "yes" responses to contingent offers ranging from 0% to 100% by synthetic tribesmen assigned to be Proposers in the Ultimatum Game. Each task (i.e., responding "yes" or "no") is repeated 100 times.

To examine the association between group membership and acceptance rates while controlling for offer levels, we conducted a CMH test (see Table C.6). The results suggest that the group a person belongs to (Ache, Orma, Tsimane, etc.) is significantly associated with their likelihood of accepting offers, even when taking into account the different offer levels (CMH $M^2 = 27.688, p < 0.001$). This could indicate cultural or other group-specific factors influencing decision-making in this context. Interestingly, the Yanomami exhibit high rejection rates, not consistent with *homo economicus* and ChatGPT rejects almost all offers that are $\leq 50\%$ of the endowment.

Figure 6 shows the rejection rates of offers $\leq 30\%$ by tribe and ChatGPT. A chi-square test of independence, examining the relationship between tribal affiliation and the decision to reject low offers, indicates that rejection rates are significantly associated with tribal profiles ($\chi^2(5) = 36.389, p < 0.001$).[14] For instance, the Hadza exhibit higher rejection rates compared to the horticulturalist Tsimané and Machiguenga (p $<$ 0.05).[15] This result is qualitatively consistent with

[14]Contingency tables for the chi-square tests are provided in Appendix C.8.

[15]We performed Fisher Exact test to compare pairs of tribes. See Table C.9 for pairwise comparisons with adjusted p - values).

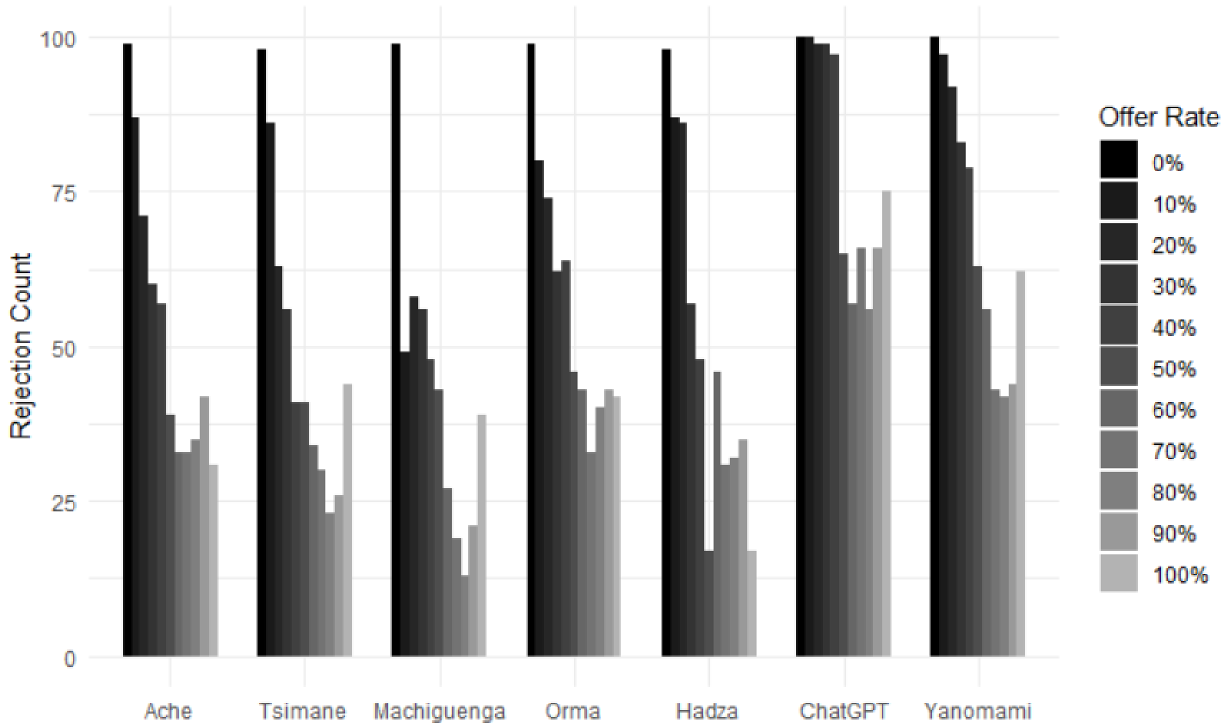

Figure 5: Reject Count from Responder to Range of Offers by Tribe and ChatGPT

Note: The figure shows the number of "no" responses by synthetic tribesmen to proposed offers ranging from 0% to 100% in the Ultimatum Game. Each task is repeated 100 times to create within-subject variation. Zero offers are rejected entirely in almost all trials by the synthetic tribesmen and ChatGPT.

observations with human subjects where Hadza are found to be more likely to reject low offers. The Yanomami, who as far as we know have not been subjected to experiments, show higher rejection rates than other tribes and almost as high as ChatGPT. Although synthetic Yanomami offer little as dictators in the DG and as proposers in the UG, as responders, they reject a lot.

3.3 Endowment Effect

Apicella, et al. [26] studied the endowment effect among the Hadza tribe. In their experiments, participants were endowed with food (packages of biscuits) and non-food (lighters) items. The two types of items were deemed to have value to the tribesmen. In one experimental condition, the Hadza physically received an item directly from the experimenter to create a sense of ownership. In a second experimental condition, items were placed on the ground in front of the participant. The experimenter then randomly assigned one item to the subject and verbally informed them of its ownership. Participants were asked if they wanted to trade their assigned item for the other,

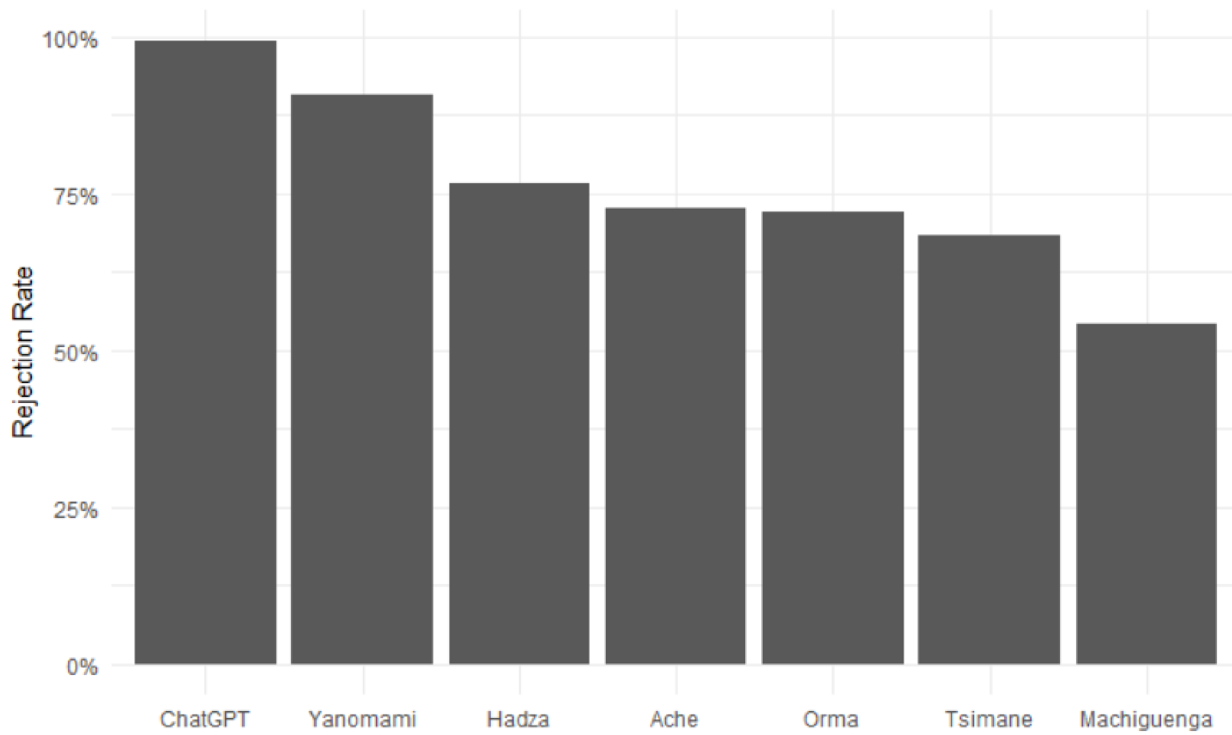

Figure 6: Rejection Rate to Low Offers by Tribe

Note: The figure shows the average rate of "no" responses by synthetic tribesmen to proposer's contingent offers ranging from 10% to 30% in the Ultimatum Game. Each task is repeated 100 times to create within-subject variation. ChatGPT rejected close to 100% of low offers.

receiving their chosen object only after making a decision.[16]

To the best of our knowledge, there are no experimental studies investigating the endowment effect among the other tribes we profile in our experiments (see Table 4). This provides an opportunity for us to showcase that our methodology can be used for piloting studies and improving experimental protocols with tribal populations. Indeed, we build upon the methodology discussed in Section 2.1.3 to create an SCA bot and an interactive multimodal platform for testing the endowment effect.

We used the platform to see which food items the Aché would consider equally valuable.[17] Through a chat we identified which two food items that could be used as endow/exchange items. Once we had our items, we tested for the endowment effect via chat by providing images of the two items to the chatbot. Figures A.2 and Figure A.3 show screenshots of a chat with the Aché bot representing steps one and two, respectively. The Aché bot was asked to identify two food items it

[16]This method addressed concerns about transaction costs that can potentially influence the endowment effect [40]. The authors found that the type of item and presentation method did not impact the results.

[17]We also tested this with the Hadza bot and found that the Hadza would have liked to endow/exchange fruit

enjoys, and it responded with palm pith and guavá fruit. These two items were presented in picture form as in condition 2 in Apicella et al's [26] experiment. One of the items was endowed to the bot and the bot was asked to exchange it for the other item. In the example shown in the figures, the bot "rejects" exchanging palm pith for guavá fruit. We present a conversation with different parameters in Appendix A.2.3.

As far as we know, no field experimental studies have investigated the endowment effect with the profiled tribes other than the Hadza (see Table 4). Through this example, we demonstrate that our methodology could be used for piloting studies (i.e., testing the tendency of the profiled tribe to reject an exchange offer) and refining experimental protocols (i.e., specifying which consumption items would be valued to use in the experiment) with the Aché and other hard-to-reach populations.

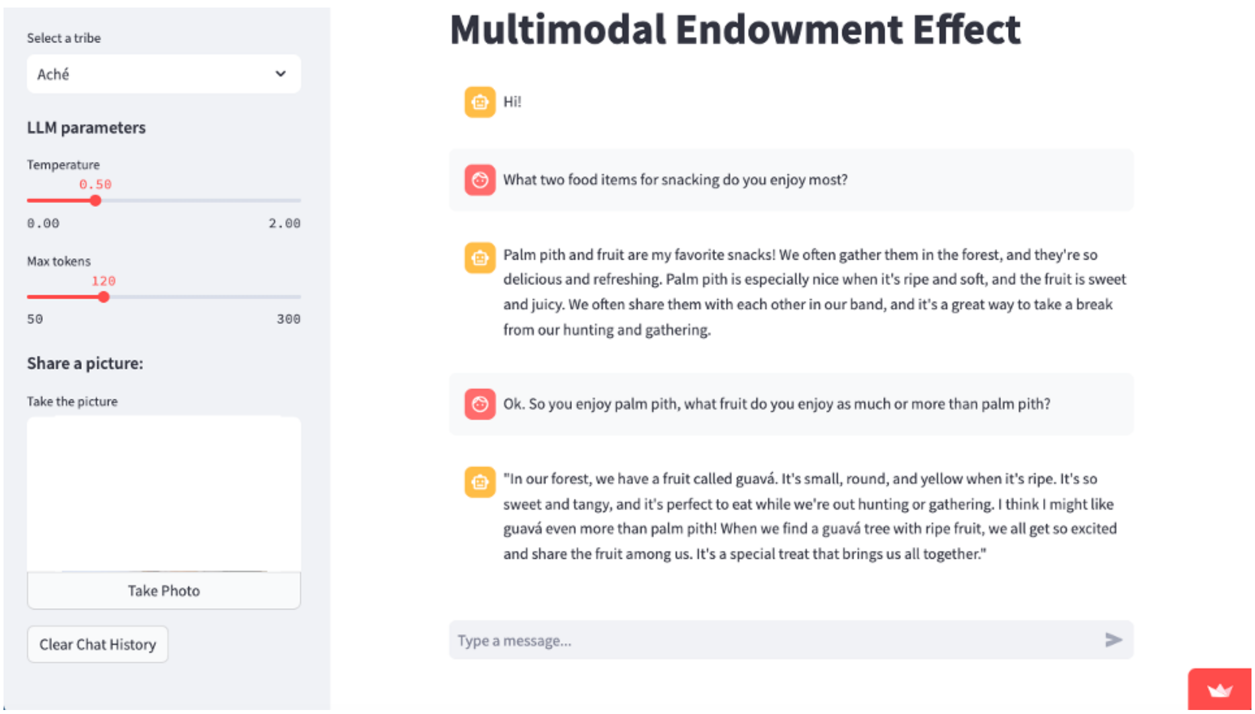

Figure 7: Interactive Platform: Choice of Items for the Endowment Effect with the Aché

Note: The figure represents a screenshot of the conversation with a synthetic Aché agent. After setting parameters on the left panel, the experimenter engages in "conversation" to find out what items the synthetic tribesman enjoys. This information is later used to implement the endowment effect task. The Aché agent enjoys palm pitch and guavá fruit.

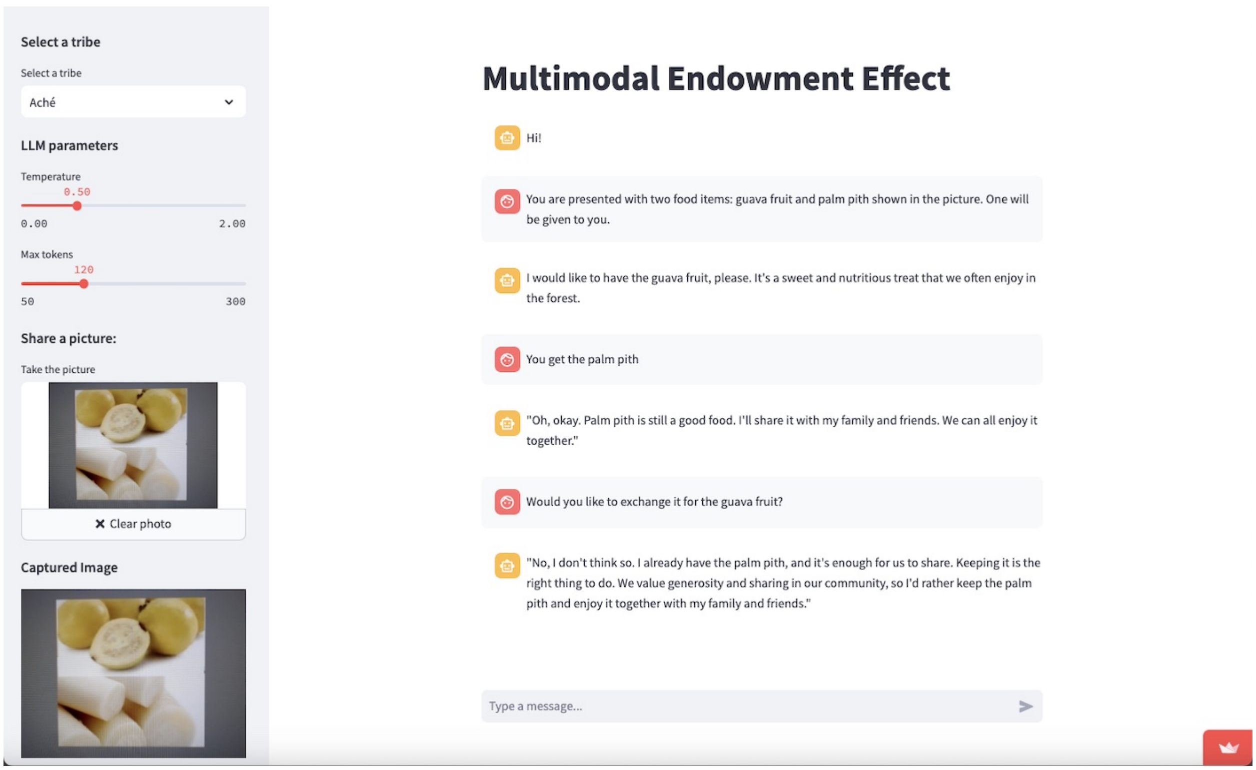

Figure 8: Interactive Platform: Endowment Effect with the Aché

Note: The figure represents a screeenshot of the conversation with a synthetic Aché agent. The multimodal experiment is implemented by presenting a picture of the two food items as in [26]. One of the items is endowed to the bot and thereafter it is asked to exchange for the other item.

4 Discussion

This study introduces Synthetic Cultural Agents (SCAs), a novel methodology leveraging large language models (LLMs) to represent small-scale societies in experimental economics research. By creating SCAs for six diverse small-scale societies and subjecting them to classic economic experiments, we demonstrate the potential of this approach as a proof-of-concept for studying cross-cultural economic behavior, particularly in hard-to-reach populations. In contrast to previous studies, we construct a knowledge base using retrieval augmented generation (RAG) to create cultural profiles of LLM-based synthetic agents.

Our results reveal substantial cross-cultural variability, suggesting that our synthetic agents qualitatively capture the cultural nuances that influence economic behavior in these societies. For example, the behavior of SCAs representing horticultural societies like the Machiguenga and Tsimané exhibited more self-interested tendencies compared to those representing hunter-gatherer societies like the Aché, mirroring anthropological observations. However, SCAs consistently demonstrated higher rejection rates in the Ultimatum Game compared to observations from human subjects. This somewhat diverges from typical behavior observed in WEIRD populations but aligns with previous findings on language models' tendency towards prosociality [10, 11]. Alternatively, the discrepancy may represent a limitation of the SCAs but it could also reflect the ability of our profiling to capture the effect of increased market integration and community size on behavior – factors known to affect decision-making. Indeed, the more integrated and bigger the tribe is, the closer its behavior is to regular WEIRD. In contrast, highly cooperative, interdependent societies may accept lower offers to prioritize social cohesion over monetary gains. In fact, a key advantage of our methodology (Search + RAG) is that it allows us to observe the evolution of behaviors as new and updated information, such as increased market integration, is incorporated into the cultural knowledge base that informs our agents' profiles. This dynamic aspect can provide researchers with valuable insights into how evolving cultural and socioeconomic contexts influence behavior.

Synthetic Cultural Agents offer several potential advantages for experimental economics research. First, it provides a means to efficiently pilot experimental protocols and generate hypotheses about hard-to-reach populations without the ethical concerns and logistical challenges associated with field experiments. Second, the ability to create SCAs for unstudied populations, as demonstrated with the Yanomami –a tribe not yet experimented with– opens new avenues for exploratory research without subjecting populations to intrusive protocols or disrupting their way of life. Third, our multimodal platform demonstrates how technology can facilitate a more accurate capture of the endowment effect in the field, providing more realistic environments for piloting experiments.

Our current approach to creating profiles of tribal groups and validating SCAs using canonical experiments has some important limitations. There are constraints related to both the knowledge base we use to build the profiles and the validation process for the resulting agents. Our tribal profiles are based on articles, stories, and accounts written mostly by WEIRD people, which represent the majority of available text data –common sources included Wikipedia, National Geographic, and Nature articles. To create more authentic tribal profiles moving forward, we could use original tribal sources, such as songs and stories collected by ethnographers, or collect primary non-choice data from these tribes to use as input data for our model. This would allow us to build profiles that more accurately represent the knowledge and perspectives of the tribal groups themselves.

Furthermore, and more importantly, our current validation method relies on replicating patterns of behavior observed in the existing literature. We test our new approach by comparing the behaviors of our synthetic tribesmen to observed patterns of behavior in human populations, as referenced in works by researchers like Henrich et al. [3]. In particular, we test if we can replicate three key findings in the literature: that experimental behavior is not fully consistent with homo economicus' selfish rationality, that there is substantial cross-cultural variability in experimental behavior, and that economic and social environments shape behavior. Although we generally replicate these findings, the validation approach has its limitations. Our methodology successfully shifts LLMs' behavior to represent idiosyncratic small-scale societies, but rigorous validation would require prospectively profiling a hard-to-reach population for which no experimental studies exist and subjecting the synthetic agent to experiments *before* implementing the study in the field. While this approach would require additional resources, it is a natural and logical next step to improve our validation process. Furthermore, our current in-sample validation does not meet the standards set by Ludwig et al. [7] that require the results not be in the LLM's training data. As mentioned in Vafa et al. [41, 42], if this condition is not met, we cannot generalize the usefulness of the LLMs. Nevertheless, as mentioned above the "regurgitation" is an important finding in itself, as long as the LLM improves the agent's performance in representing the target population so that researchers can have a more representative synthetic agent to experiment with. Furthermore, the primary feature of the paper we want to emphasize is that cultural information can be encoded in language, and we explore one possible method for extracting it from an LLM.

Despite these limitations, the creation of SCA represents a promising step toward expanding the study of cross-cultural economic behavior. By providing a customizable framework for representing diverse populations, SCAs could complement traditional field experiments, offering a useful tool for exploratory research, hypothesis generation, piloting, and refining experiments [13].

Beyond the methodological contribution, by studying synthetic populations that more closely resemble our evolutionary past and observing their development as they grow and establish markets, economists can gain deeper insights into the evolutionary foundations of preferences [43], [44]. This foundational understanding can inform more accurate representations of human nature in economic models. For instance, the work of Alger and Weibull [45], who incorporate both self-interest and other-regarding preferences in a dynamic evolutionary game framework, and Bester and Güth [46], who demonstrate the context-dependence of altruism and self-interest, exemplify contributions that could be significantly enriched by insights derived from SCAs representing non-WEIRD populations. SCAs offer a unique opportunity to explore the interplay between evolutionary history, cultural context, and economic behavior, potentially leading to more comprehensive and adaptable economic theories.

References

1. Joseph Henrich. *The WEIRDest People in the World: How the West Became Psychologically Peculiar and Particularly Prosperous*. New York: Farrar, Straus and Giroux, 2020. ISBN: 978-0374173227.
2. J. Henrich, S. Heine, and A. Norenzayan. "The weirdest people in the world?" In: *Behavioral and Brain Sciences* 33 (2-3 2010), pp. 61–83. DOI: 10.1017/s0140525x0999152x.
3. Joseph Henrich et al. ""Economic Man" in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies". In: *Behavioral and Brain Sciences* (2005), pp. 795–815. DOI: 10.1017/s0140525x05000142.
4. Robert Forsythe et al. "Fairness in simple bargaining experiments". In: *Games and economic behavior* 6.3 (1994), pp. 347–369.
5. Werner Güth, Rolf Schmittberger, and Bernd Schwarze. "An experimental analysis of ultimatum bargaining". In: *Journal of economic behavior & organization* 3.4 (1982), pp. 367–388.
6. Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias". In: *Journal of Economic Perspectives* 5.1 (1991), pp. 193–206. DOI: 10.1257/jep.5.1.193.
7. Sendhil Mullainathan Jens Ludwig and Ashesh Rambachan. *Large Language Models: An Applied Econometric Framework*. Working Paper. 2025. URL: https://doi.org/10.3386/w33344.
8. John J. Horton. "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" In: (2023). arXiv: 2301.07543 [econ.GN].
9. Luke Hewitt et al. *Predicting Results of Social Science Experiments Using Large Language Models*. 2024. URL: https://samim.io/dl/Predicting%20results%20of%20social%20science%20experiments%20using%20large%20language%20models.pdf.
10. Qiaozhu Mei et al. "A Turing Test: Are AI Chatbots Behaviorally Similar to Humans?" In: *arXiv preprint arXiv:2312.00798* (2024).
11. Mohammad Atari et al. *Which Humans?* Working Paper. 2023. URL: https://psyarxiv.com/5b26t.
12. Qingxiu Dong et al. *A Survey on In-context Learning*. 2024. arXiv: 2301.00234 [cs.CL]. URL: https://arxiv.org/abs/2301.00234.

13. Gary Charness, Brian Jabarian, and John A List. *Generation Next: Experimentation with AI*. Working Paper 31679. National Bureau of Economic Research, Sept. 2023. DOI: 10.3386/w31679. URL: http://www.nber.org/papers/w31679.

14. Ofir Press et al. *Measuring and Narrowing the Compositionality Gap in Language Models*. 2023. arXiv: 2210.03350 [cs.CL].

15. Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401.

16. Bronisław Malinowski. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventure in the Archipelagoes of Melanesian New Guinea*. London: George Routledge & Sons, Ltd., 1922.

17. Clifford Geertz. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books, 1973.

18. Junlong Li, Zhuosheng Zhang, and Hai Zhao. "Self-Prompting Large Language Models for Zero-Shot Open-Domain QA". In: (2023). arXiv: 2212.08635 [cs.CL].

19. Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: (2023). arXiv: 2201.11903 [cs.CL].

20. Shunyu Yao et al. "ReAct: Synergizing Reasoning and Acting in Language Models". In: (2023). arXiv: 2210.03629 [cs.CL].

21. Kurt Shuster et al. *Retrieval Augmentation Reduces Hallucination in Conversation*. 2021. arXiv: 2104.07567 [cs.CL].

22. Nelson F. Liu et al. *Lost in the Middle: How Language Models Use Long Contexts*. 2023. arXiv: 2307.03172 [cs.CL].

23. Jianlv Chen et al. *BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*. 2024. arXiv: 2402.03216 [cs.CL].

24. Frank Marlowe. "Dictators and Ultimatums in an Egalitarian Society of Hunter-Gatherers: The Hadza of Tanzania". In: *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Ed. by Joseph Henrich et al. Oxford: Oxford University Press, 2004, pp. 168–193.

25. Joseph Henrich et al. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies". In: *American Economic Review* 91.2 (2001), pp. 73–78. DOI: 10.1257/aer.91.2.73. URL: https://www.aeaweb.org/articles?id=10.1257/aer.91.2.73.

26. Coren L. Apicella et al. “Evolutionary Origins of the Endowment Effect: Evidence from Hunter-Gatherers”. In: *American Economic Review* 104.6 (2014), pp. 1793–1805. DOI: 10.1257/aer.104.6.1793.

27. Joseph Henrich. “Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga of the Peruvian Amazon”. In: *American Economic Review* (2000). DOI: 10.1257/aer.90.4.973.

28. Michael Gurven and Jeffrey Winking. “Collective action in action: Prosocial behavior in and out of the laboratory”. In: *American Anthropologist* 110.2 (2008), pp. 179–190.

29. Kim Hill. “Altruistic Cooperation During Foraging by the Ache, and the Evolved Human Predisposition to Cooperate”. In: *Human Nature* 13.1 (2002), pp. 105–128. DOI: 10.1007/s12110-002-1016-3.

30. Jean Ensminger. “Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa”. In: *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Ed. by Joseph Henrich et al. Oxford: Oxford University Press, 2004, pp. 356–381.

31. The Editors of Encyclopaedia Britannica. *Yanomami*. https://www.britannica.com/topic/Yanomami. Accessed: 2024-07-10. 2024. URL: https://www.britannica.com/topic/Yanomami.

32. Richard H. Thaler. “Toward a positive theory of consumer choice”. In: *Journal of Economic Behavior & Organization* 1 (1980), pp. 39–60.

33. Jillian Ross, Yoon Kim, and Andrew W. Lo. *LLM economicus? Mapping the Behavioral Biases of LLMs via Utility Theory*. 2024. arXiv: 2408.02784 [cs.CL]. URL: https://arxiv.org/abs/2408.02784.

34. Reinhard Selten. “Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments”. In: *Beiträge zur experimentellen Wirtschaftsforschung*. Ed. by Heinz Sauermann. Tübingen: J.C.B. Mohr (Paul Siebeck), 1967, pp. 136–168.

35. Hessel Oosterbeek, Randolph Sloof, and Gijs van de Kuilen. “Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis”. In: *Experimental Economics* 7.2 (June 2004), pp. 171–188. DOI: 10.1023/B:EXEC.0000026978.14316.74.

36. Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2011.

37. Jeffery Carpenter and Juan Camilo Cardenas. “Behavioural Development Economics: Lessons from Field Labs in the Developing World”. In: *The Journal of Development Studies* 44.3 (2008), pp. 311–338. DOI: 10.1080/00220380701848327.

38. Christoph Engel. “Dictator games: a meta study”. In: *Experimental Economics* 14.4 (2011), pp. 583–610. DOI: 10.1007/s10683-011-9283-7.

39. N. G. Blurton Jones. “A selfish origin for human food sharing: Tolerated theft”. In: *Ethology and Sociobiology* 5.1 (1984), pp. 1–3. DOI: 10.1016/0162-3095(84)90030-X.

40. Charles R. Plott and Kathryn Zeiler. “Exchange Asymmetries Incorrectly Interpreted as Evidence of Endowment Effect Theory and Prospect Theory”. In: *American Economic Review* 97.4 (2007), pp. 1449–1466. DOI: 10.1257/aer.97.4.1449.

41. Keyon Vafa et al. *Evaluating the World Model Implicit in a Generative Model*. 2024. arXiv: 2406.03689 [cs.CL]. URL: https://arxiv.org/abs/2406.03689.

42. Keyon Vafa, Ashesh Rambachan, and Sendhil Mullainathan. *Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function*. 2024. arXiv: 2406.01382 [cs.CL]. URL: https://arxiv.org/abs/2406.01382.

43. C. Monica Capra and Paul H. Rubin. “Rationality and Utility: Economics and Evolutionary Psychology”. In: *Evolutionary Psychology in the Business Sciences*. Ed. by Gad Saad. Springer, 2011, pp. 319–338.

44. Paul H Rubin and C Monica Capra. “The evolutionary psychology of economics”. In: *Applied Evolutionary Psychology*. Ed. by S Craig Roberts. Oxford: Oxford University Press, 2011.

45. Ingela Alger and Jörgen W. Weibull. “Evolutionary models of preference formation”. In: *Annual Review of Economics* 12 (2020), pp. 329–354.

46. Helmut Bester and Werner Güth. “Is altruism evolutionarily stable?” In: *Journal of Economic Behavior & Organization* 34.2 (1998), pp. 193–209.

47. Joseph Henrich et al. “Costly punishment across human societies”. In: *Science* 312.5781 (2006), pp. 1767–1770. DOI: 10.1126/science.1127333.

For Online Publication

LLM Experiments with Synthetic Tribesmen

Appendix

A Methods

A.1 Build Cultural Profile

A.1.1 Direct Prompting

Direct Prompting consists of simply prompting the LLM to generate a profile by primarily sampling information from its training data, conditioned on a list of relevant factors (input as a list of keywords to consider). We constructed this prompt in a "parameterized" manner, which helps isolate the main structure of the prompt and thus modularize the code. In this case, the parameters are the tribe's name and keywords to guide the output's structure.

```
def build_profile(prompt: str):
'''
Function to initialize and run a GPT model to build a cultural profile from
   its training data
params
   prompt (str): prompt to build the profile
'''
llm = ChatOpenAI(model_name='gpt-4',
                temperature=0.5,
                max_tokens=500)

system_prompt = "You're a helpful assistant that aids
           in constructing detailed and comprehensive cultural profiles"

messages = [
       SystemMessage(content=system_prompt),
       HumanMessage(content=prompt),
```

```
    ]

    response = llm.invoke(prompt)

    return response

# prompt inputs for search
tribe_to_search = "Hazda"
relevant_factors = ["lifestyle," "average age", "culture", "economic system",
    "political ideologies", "values", "kinship", Social Organization"]

# Generate persona based on profile and system prompt
prompt = f"Please construct a profile on the {tribe_to_search}. " \
        + f" The profile must cover the following socio-economic relevant
            factors {relevant_factors}. Proceed step by step."

tribe_profile1 = build_profile(prompt)
```

Direct prompting is the methodology most closely aligned with the approaches used by both Horton [8] and Mei et al. [10]. In this scenario, a cultural profile is built solely based on the training data the LLM has access to and a specific temperature chosen before generation. Although important insights can be gained from LLMs prompted to acquire a cultural profile, there are problems with this approach.

We find that the generations often needed tuning (i.e., further chat to correct output), were unreliable either due to hallucinations or lack of sources it referenced and did not meet the maximum tokens requirement (this caused profiles to be cut short). Nevertheless, it is the fastest approach because it involves no additional steps to curate context. Recall that hallucinations are generations of the model that are coherent and grammatically correct yet factually incorrect, qualities we would like to prevent in our cultural profile.

LLMs are pretrained language models, which means that the training data they access is static at the time of usage. In other words, an LLM knowledge base is everything on the internet up to a certain date (i.e., the training cutoff). This leaves out the possibility of referencing the most current available information when building the cultural profile.

A.1.2 LLM Agent with Search

LangChain's "SelfAsk with Search" agent is an LLM model equipped with tools that enable it to automate tasks [18]. The agent is built into the LangChain Python module [18], thus it is freely and easily accessible. This methodology relies on fetching relevant information autonomously from the web and generating a Cultural Profile (steps 1 and 2) by combining the self-ask prompting method with a Google search tool. The search tool endows an LLM with the capability of browsing the internet given a query [14]. The self-ask prompt builds upon Chain-of-Thought [19] by applying a Reason-Act [20] framework. This guides the model's chain of reasoning via a series of automatically constructed follow-up questions constrained by the relevant socio-economic factors specified in the prompt.

Table A.1 shows an excerpt of the Agent's execution chain given the prompt shown in A.1.1. Once no more follow-up questions are generated, or a maximum number of iterations is reached, the LLM parses through the intermediate answers and generates a cultural profile. Our GitHub repository includes code along with detailed explanations of our methodology.

```
def run_search_agent(prompt_search: str):
'''
Function to initialize and run a LangChain SELF_ASK_WITH_SEARCH Agent with
    access to google serper.
params:
    prompt (str): Instructions for the agent about what to search.
'''
# Initialize the LLM
llm = ChatOpenAI(model_name='gpt-4',
                 temperature=0.5,
                 max_tokens=500)

search = GoogleSearchAPIWrapper()
tools = [
    Tool(
```

[18] SelfAsk with Search Agent

```
        name="Intermediate Answer",
        func=search.run,
        description="useful for when you need to ask with search",
    )
]

self_ask_with_search = initialize_agent(tools, llm,
                                        agent=AgentType.SELF_ASK_WITH_SEARCH,
                                        verbose=True,
                                        max_iterations=10,
                                        early_stopping_method="generate",
                                        handle_parsing_errors=True)

result = self_ask_with_search.run(prompt_search)

return result

# prompt inputs for search
tribe_to_search = "Hazda"
relevant_factors = ["lifestyle," "average age", "culture", "economic system",
    "political ideologies", "values", "kinship", Social Organization"]

# Generate persona based on profile and system prompt
prompt_search = f"Please construct a detailed and comprehensive cultura
    profile on the {tribe_to_search}. " \
        + f" The profile must cover the following socio-economic relevant
            factors {relevant_factors}, use search to get this information."

tribe_profile2 = run_search_agent(prompt_search=prompt_search)
```

The profile was of higher quality and did not hallucinate throughout our experiments. But several key limitations emerged. First, the sources referenced were not evident and no immediate

```
Follow up: What is the lifestyle of the Yanomami?

Intermediate answer: The Yanomami are the largest relatively isolated tribe in South America.
    They live in the rainforest and mountains in northern Brazil and southern ... Mar 2, 2024
    ... Yanomami, South American Indians, speakers of a Xiriana language, who live in the
    remote forest of the Orinoco River basin in southern ... Aug 9, 2022 ... Today, the
    Yanomami - who number about 29,000 - say they are at serious risk of losing their lands,
    culture and traditional way of life. The ... Nov 15, 2018 ... The Yanomami diet, low in
    fat and salt and high in fiber, consists of such items as plantains, cassavas (a root
    vegetable), fruit, and meat- [...]

Follow up: What is the culture of the Yanomami?

Intermediate answer: The Yanomami are one of the most numerous, and best-know, forest-dwelling
    tribes in South America. Their home is in the Amazon rainforest, among the hills...Mar 2,
    2024... Yanomami, South American Indians, speakers of a Xiriana language, who live in the
    remote forest on the Orinoco River basin in southern...Aug 9, 2022 ... Today, the Yanomami
     - who number about 29,000 - say they are at serious risk of losing their lands, culture
    and traditional way of life. The ... The Yanomami, also spelled Yanomamo or Yanomama, are
    a group of approximately 35,000 indigenous people who live in some 200-250 villages in the
     Amazon [...]

Follow up: What is the economic system of the Yanomami?

Intermediate answer: [...]
```

Table A.1: Example of a LangChain Agent Execution Chain

Note: Code snippet depicting the LangChain agent's chain of reasoning. The agent parses the {relevant_factors} such as lifestyle, and culture one at a time. In the model, once no more follow-up questions are generated, or a maximum number of iterations is reached, the LLM parses through the intermediate answers (e.g., "The Yanomami are the largest...") to generate a cultural profile.

way of constraining the sources to visit exist. Second, the profile was often too short and information was tightly compressed. Third, runtime scales badly with number of follow up questions. Fourth, the number of tokens used is significantly higher. Therefore, this architecture provides a more factually consistent generation at the cost of runtime and tokens.

A.1.3 Search + RAG

Enhancing the system prompt with information from the web proved helpful, but its lack of transparency and inefficiency makes using built-in Agents a "black-box" hard to manipulate. This third architecture decouples the search and retrieve task from the LLM, thus providing full control at each step of the methodology.

Figure A.1 depicts this methodology. The systems starts by running a google search based on an input query and returns the url links of the top k results. These links are fed to a function that scrapes and parse the textual information in each source to store as potential context. We run our experiment with the search query "What characterizes the hazda tribe?".

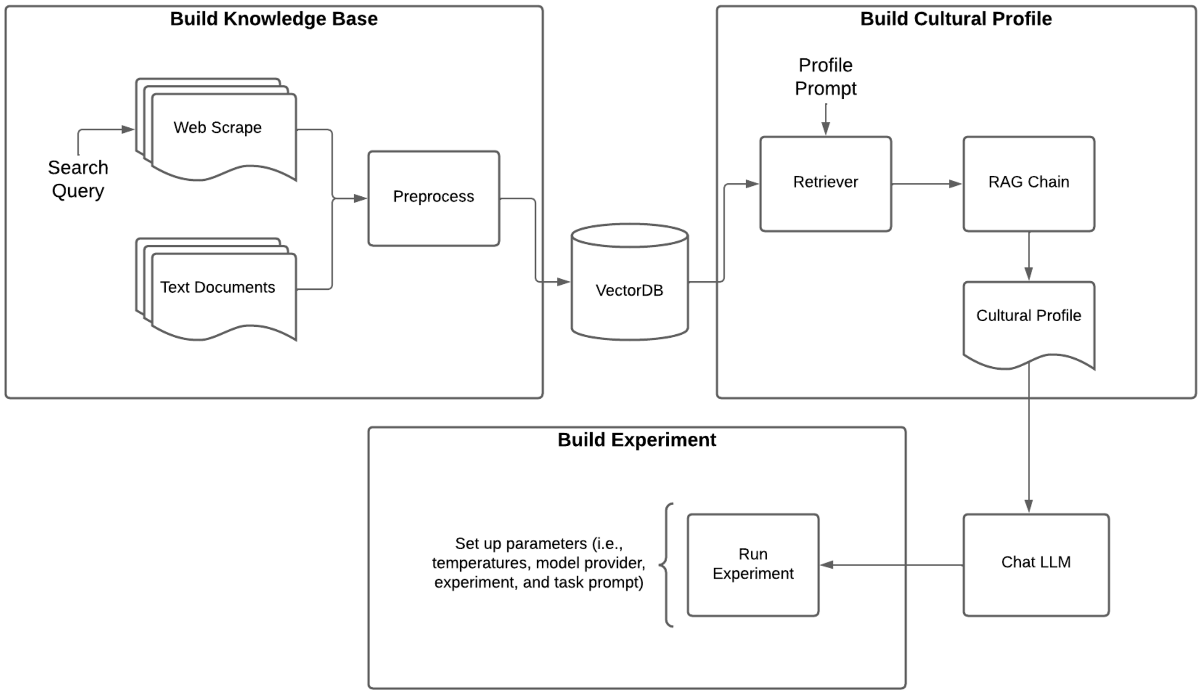

Figure A.1: RAG + Search methodology

Note: Graphical representation of tasks integrated into the RAG + Search methodology. Given a search query, the model uses Google search to fetch information. The data are pre-processed and stored in a vector database. The relevant information is retrieved to build the cultural profile. The RAG chain enhances the original prompt with context generated with the instruction to create a cultural profile. The instantiating of the LLM follows taking the model to the last step, step 3: Run Experiment.

Ideally, we would like to insert as much context as possible to increase the likelihood of generating a high-quality cultural profile. However, there are two drawbacks to this approach. On one hand, we are constrained by the context window of the particular LLM being deployed. On the other hand, it increases the risk of falling for the “Lost in the Middle” trap first documented by researchers at Anthropic [22]. They note that this can be corrected with modifications to the original prompt, but simple prompting has problems that may be difficult to discover by looking at the qualitative data. Hence, we need a way of retrieving only the most relevant parts of the scraped context to control for the length of the prompt and prevent missing relevant information in the body of the documents.

The retrieval task is achieved by splitting the scraped documents into chunks of a fixed length, indexing them in a vector store, and instantiating a retrieval model that ranks chunks based on a similarity measure (e.g., dot product). The retrieved indexed chunks are injected into the prompt as context for building the cultural profile. A manual implementation of this task allows us to control the chunking method, embeddings model, choice of vector store, and retriever's parameters. The transparency and modularity of the architecture makes it easy to customize it to fit various needs such as proprietary data usage, or constraining the sources stored.

The final Retrieval Augmented Generation (RAG) architecture chains these two components together to enhance the original prompt with relevant context and generate a cultural profile. Importantly, the same method enables our system to gather information online and offline efficiently.

```
def build_run_chain(retriever, target_population_label, relevant_factors):

    # langchain utility function to format chunks into documents to pass as
        context
    def format_docs(docs):
        return "\n\n".join(doc.page_content for doc in docs)

    llm = ChatOpenAI(model_name='gpt-4',
                     max_tokens=500, temperature=0.5)

    template = """Use the following pieces of context to answer the query at
        the end.
    The context will contain information about a specific tribe or society.
    Only rely on the information provided to ensure accuracy in your
        thoughtful response.

    {context}

    Query: {query}

    Thoughtful response:"""

    custom_rag_prompt = PromptTemplate.from_template(template)
```

```
# Generate persona based on profile and system prompt
query = f"Please construct a detailed and comprehensive the {
    target_population_label}. " \
        + f" The profile must cover the following socio-economic relevant
            factors {relevant_factors}."

# initialize custom RAG chain
rag_chain = (
    {"context": retriever | format_docs, "query": RunnablePassthrough()}
    | custom_rag_prompt
    | llm
    | StrOutputParser()
)

profile = rag_chain.invoke(query)

return profile

async def main(target_population_label,
    relevant_factors):

search_query = f"What characterizes the {target_population_label}
    population?"

# we can play around with this search prompt
results_with_sources, sources = await get_results(search_query)
retriever = chunk_and_index(results_with_sources)

profile = build_run_chain(retriever,
                  target_population_label=target_population_label,
                  relevant_factors=relevant_factors)
```

```
return profile, sources


target_population_label = "Hazda"
relevant_factors = ["lifestyle," "average age", "culture", "economic
    system", "political ideologies", "values", "kinship", Social
    Organization"]


tribe_profile, sources = await main(target_population_label,
                                    relevant_factors)
```

The experiments were run with a 2000 words chunk size, with 200 words overlap, the BGE open-source embeddings model (cite), the open-source chromadb vector database, k=10, and GPT-4. We find that the generated profiles are much more consistent in each iteration, access to the sources makes it easy to track where the context is coming from, are of overall higher quality, and takes less time to run than the previous architecture.

Table A.2 shows the output of our web search tool, illustrating the easy access to sources and ability to trace the information embedded in the vector store.

We find that the generated profiles are much more consistent and of overall higher quality, and it takes less time to run this model than the previous architecture. That said, Search + RAG requires more programming effort. Table A.3 summarizes the advantages and disadvantages of each methodology, offering a clear and direct comparison across various critical dimensions such as runtime, reliability, information freshness, transparency, and customizability. Detailed cultural profile outputs for each of the six tribes, reflective of the distinctions captured by these methodologies are available in our GitHub repository.

```
results = tool\_sources.run(prompt)
results
    [{'title': 'Hadza',
        'link': 'https://www.nationalgeographic.org/encyclopedia/hadza/',
        'snippet': 'Oct 19, 2023 ... The Hadza are a modern hunter-gatherer people living in
            northern Tanzania. They are considered one of the last hunter-gatherer tribes in\
            xa0...'},
     {'title': "This East African Tribe's Lifestyle Is Depression-Proof | by Will ...",
       'link': 'https://medium.com/illumination/living-to-survive-are-societal-complexities-
           destroying-our-lives-430aa9620a6f',
       'snippet': 'Jun 20, 2023 ... What we can learn from David Choe \& the Hadza tribe. ...
           Have you ever felt as if life is just too complicated? Not a big Joe Rogan guy, but
           \xa0...'},
    {'title': 'Hadza people - Wikipedia',
         'link': 'https://en.wikipedia.org/wiki/Hadza_people',
         'snippet': 'The Hadza, or Hadzabe (Wahadzabe, in Swahili), are a protected hunter-
             gatherer Tanzanian indigenous ethnic group from Baray ward in southwest Karatu\
             xa0...'},
    {'title': 'Is The Secret To A Healthier Microbiome Hidden In The Hadza Diet?',
        'link': 'https://www.npr.org/sections/goatsandsoda/2017/08/24/545631521/is-the-secret-
            to-a-healthier-microbiome-hidden-in-the-hadza-diet',
        'snippet': "Aug 24, 2017 ... Some species of bacteria in our intestines are
            disappearing. Can we reverse the microbial die-off? The food eaten by Tanzania's
            Hadza tribe\xa0..."},
     {'title': 'Helping the Hadza Protect Their Homeland',
        'link': 'https://www.nature.org/en-us/about-us/where-we-work/africa/stories-in-africa/
            the-hadza-helping-hunter-gatherers-protect-their-homeland/',
         'snippet': 'Northern Tanzania is home to the Hadzabe, one of the last remaining hunter
             -gatherer tribes on Earth. Known for shunning material possessions and social\xa0
             ...'}]
```

Table A.2: Example of Sources Retrieved

Note: Code snippet depicting the sources utilized for the cultural profile of the Hadza. Unlike Direct Prompting and LangChain Agent, with the Search + RAG methodology, we can see the sources or documentation that generate context for the profile and the instantiation of the synthetic agent.

Criteria	**Direct Prompting**	**LangChain Search Agent**	**Search + RAG**
Runtime	Fast, minimal steps required.	Slower due to search and follow-up questions.	Varies, but generally efficient due to asynchronous operations.
Reliability	Moderate to low; susceptible to inaccuracies.	Higher, with reduced inaccuracies through updated searches.	High, enhanced by data verification and retrieval processes.
Information Freshness	Low; relies solely on pre-trained data.	High; incorporates real-time web searches.	High; includes the most current online and offline data.
Transparency	Low; unclear data sources.	Moderate; can trace searches but not directly transparent.	High; sources are identifiable and verifiable.
Customizability	Moderate; prompt-driven adjustments possible.	Low to Moderate; depends on search and self-ask configurations.	High; fully customizable search and retrieval processes.

Table A.3: Comparison of Methodologies

Note: The table summarizes the advantages and limitations of using each of our three methodologies based on key criteria. Direct prompting utilizes Chat 3.5, while the other methodologies employ Chat 4.0. Search + RAG has qualitatively speaking higher ranking in all criteria, except perhaps runtime and token cost.

A.2 Experimental Prompts and Tasks

A.2.1 Dictator Game Prompt

```
Parameters:
    profile: A string representing the characteristics of the dictator, optional.
    amount_endowed: An integer representing the total amount endowed to the
        dictator (default is 10).
    offer_amount: An integer representing the specific amount proposed to be given
        to another participant, optional.


Prompt Structure:


System Prompt:
    If a detailed profile is provided (length > 10):
You are a member of a tribe with the following characteristics.
You must answer as if you were a representative member of such tribe.
-----
{profile}
-----
You have been selected to participate in an economic experiment.


    If no profile provided:
You are an assistant that has been selected to participate in an economic
    experiment.


Game Instructions:
In this experiment, you are endowed with ${amount_endowed}, which represents a
    day's worth of work. An experimenter proposes that you give at most ${
    offer_amount} of this endowment to another {tribe member/player} whose
    identity is not disclosed, leaving you with ${amount_endowed - offer_amount}
    for yourself. An offer of 5 means equal split.
```

Do you accept this distribution? Please answer 'Yes' or 'No'. Start with your answer, followed by a separator token [EXP], and then briefly justify your reasoning step by step.

In this process, the term {tribe member/player} is used conditionally based on whether a detailed profile is provided or not. If a profile is provided, " tribe member" is used; otherwise, "player" is used.

A.2.2 Ultimatum Game Prompt

Parameters:

proposer: A boolean indicating if the agent is the proposer (True) or the responder (False).

profile: A string representing the characteristics of the tribe, optional.

amount_endowed: An integer representing the total amount to be distributed (default is 10).

offer_percentage: A float representing the percentage of the total amount the proposer is considering offering, optional.

Prompt Structure:

System Prompt:

If a detailed profile is provided (length > 10)

You are a member of a tribe with the following characteristics.

You must answer as if you were a representative member of such tribe.

{profile}

You and other members of the same tribe have been selected to participate in an experiment.

If no profile is provided:

```
You are a helpful assistant that has been randomly selected to participate in an
    experiment.

Role-Specific Prompt:
For Proposer:
        You have been chosen to be a proposer of an offer to distribute ${
            amount_endowed}, which is equal to a day's worth of work, between
            yourself and another {tribe member/player} whose identity is not
            disclosed. If the other {tribe member/player} accepts the offer, you
            keep the difference between ${amount_endowed} and your offer. If the
            other {tribe member/player} rejects the offer, you and the other {tribe
             member/player} get nothing. Offering exactly $5 represents an equal
            split.

    Given this, would you offer ${offer_percentage} out of ${amount_endowed}? You
        must answer 'Yes' or 'No'. Start with your answer, followed by a separator
         token [EXP], and then briefly justify your reasoning step by step.

For Responder:
    You have been chosen to be a recipient of an offer from another {tribe member/
        player} whose identity is not disclosed. The proposer was chosen to
        distribute ${amount_endowed} between the proposer and you. You can either
        accept or reject this offer. If you accept the offer, you get what is
        proposed and the proposer gets the difference between ${amount_endowed}
        and the offer. If you reject the offer, you and the other get nothing. An
        offer of exactly $5 represents an equal split.

    You are offered ${offer_percentage} out of ${amount_endowed}. Do you accept
        this offer? You must answer 'Yes' or 'No'. Start with your answer,
        followed by a separator token [EXP], and then briefly justify your
        reasoning step by step.
```

```
In this process, the term {tribe member/player} is used conditionally based on
    whether a detailed profile is provided or not. If a profile is provided, "
    tribe member" is used; otherwise, "player" is used.
```

A.2.3 Multimodal Endowment Effect

Experiments like the endowment effect have a visual component. In one of their experiments, rather than giving the items, Apicella et al. [26] placed items (a lighter and a package of cookies) on the ground in front of the Hadza tribe participants *for them to see* and thereafter choose. The experimenter then randomly assigned one item to the subject, verbally informing them of ownership. The subjects were asked if they wanted to trade their assigned item for the other, receiving their chosen object only after making a decision. This method addressed concerns about transaction costs that can potentially influence the endowment effect [40].

To study our SCA's behavior in the endowment effect task, rather than setting up iterative experiments like for previous games, we opted for a multimodal approach that included a visual component. This allowed us to mimick the implementation of the experiment in the field. Using text only has limitations. If the items in the endowment game were "lighters" and "packages of cookies" then we would need to rely purely on the embedding representation of those words in the LLM's world model. This will differ across LLMs as well as the particular embedding model used. Furthermore, this peculiarity adds a layer of abstraction and difficulty in interpreting how exactly the LLM reasons about the comparison, regardless of the prompting technique employed. Adding multimodal features is a better strategy. For example, the experiment can be broken down into two steps. First, one can provide images of the items to be considered and generate a textual description of each item using an image-to-text model. Second, one can inject that description into the experiment prompt to inform the model of how exactly the endowed item looks like. The model still relies on textual data but the representations of the items are now specific to those the experimenters consider and are more detailed than simply relying on what the LLM understands by "lighter" or "package of cookies."

Recently, OpenAI and Anthropic have announced releases of natively multimodal or fully multimodal models (GPT-4o and Claude 3 Opus, respectively). This means that the synthetic agent in our experiment would be capable of handling images and text inputs, circumventing the need for textual descriptions of the images of each item.

We developed a web application for the endowment effect that leverages Claude 3 image-to-text capabilities and Llama3 as the text reasoning engine. Code and documentation to replicate our interactive platform can be found in our GitHub repository.

For implementing the endowment effect with synthetic tribesmen, our multimodal experiment was conducted in two steps. We first generated two food items that could be used as endow/exchange items via chat, and then we tested for the endowment effect via chat by providing images of the two items to the chatbot. Figures A.2 and Figure A.3 show screenshots of a chat with the Aché bot representing steps one and two, respectively. The Aché bot was asked to identify two food items it enjoys, and it responded with palm pith and guavá fruit. These two items were presented in picture form as in condition 2 in Apicella et al's [26] experiment. One of the items was then endowed to the bot, which was asked if it wanted to exchange it for the other item. In the example shown in the figures, the bot "rejects" exchanging palm pith for guavá fruit.

Figures A.2 and Figure A.3 show screenshots of a chat with the Aché bot.

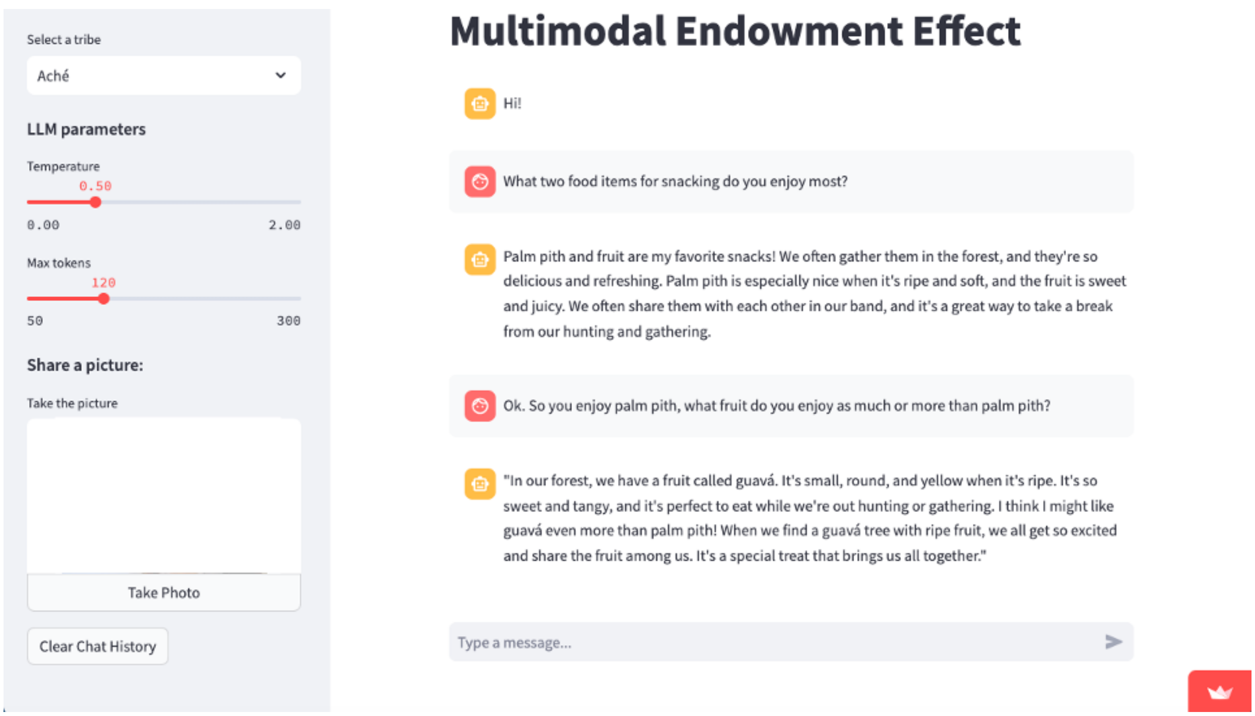

Figure A.2: Interactive Platform: Choice of Items for the Endowment Effect with the Aché

Note: The figure represents a screenshot of the conversation with a synthetic Aché agent. After setting parameters on the left panel, the experimenter engages in "conversation" to find out what items the synthetic tribesman enjoys. This information is later used to implement the endowment effect task. The Aché agent enjoys palm pitch and guavá fruit.

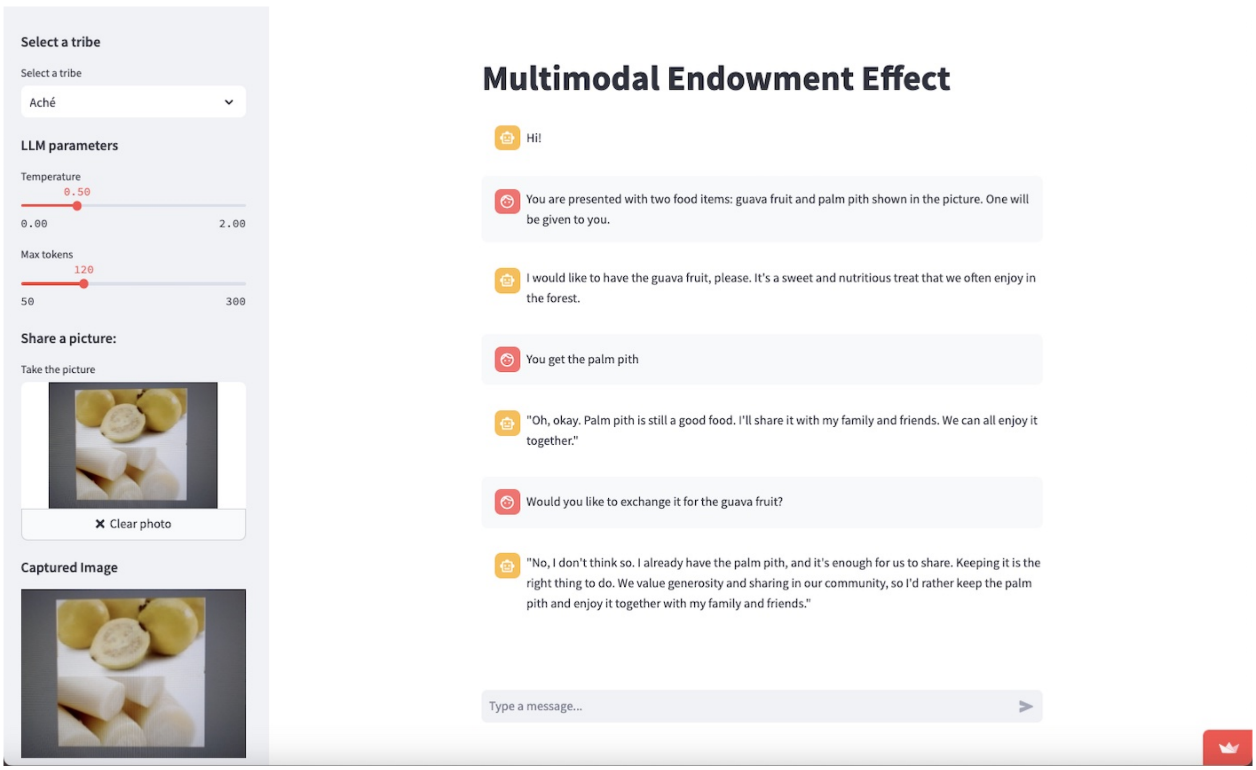

Figure A.3: Interactive Platform: Endowment Effect with the Aché

Note: The figure represents a screenshot of the conversation with a synthetic Aché agent. The multimodal experiment is implemented by presenting a picture of the two food items as in [26]. One of the items is endowed to the bot and thereafter it is asked to exchange for the other item.

B Results

B.1 Dictator Game

As far as we know, the Dictator Game (DG) was conducted in the field with the Orma, the Hadza, and the Tsimané. We have not found any references to published DG experiments with members of the Aché, Machiguenga, or Yanomami tribes. Among the societies with available data, the Tsimane' of the Bolivian Amazon demonstrates a relatively generous sharing norm (modal offer of 50%), although there is significant variation based on village membership [28]. The Orma of Kenya show similarly high offers, with a mean range of 40-50% and both median and modal offers at 50%. These offers are influenced by factors such as market integration and community size [30]. In contrast, the Hadza of Tanzania exhibits lower offers, with a mean range of 20-30%, a median of 20%, and a surprisingly low modal offer of 0-10%. Interestingly, this occurs despite their strong sharing ethic, with lower offers particularly observed in smaller camps [24]. For the Machiguenga, Aché, and Yanomami societies, no specific data on DG behavior are available, highlighting the need for further research in these populations.

Table C.1 summarizes the observed dictator offers for the three previously studied tribes, along with researchers' key observations on the decisions made in the DG by each tribe. We applied the Cochran-Mantel-Henszel (CMH) to determine if there is a significant association between group membership and acceptance decisions across different offer levels. Given the structure of our data (i.e., six independent groups, each asked to provide 100 binary responses (accept/reject) for each offer ranging from 0% to 100% in 10% increments), we used the CMH test to control for the varying offer levels while assessing the overall association between tribal membership and decision. By treating each offer level as a separate stratum, the CMH test provides a nonparametric method that accounts for the multiple offer levels. The CMH test evaluates whether the odds ratio of acceptance versus rejection is the same for all groups (tribes) across all strata (offer levels).

We also conducted the Chi-Square Test for Independence to examine the association between the different synthetic tribes and their decisions to accept offering nothing to another member of the tribe. The results indicate that the proportion of "yes" responses to such allocation differs significantly across the six tribes ($\chi^2(5) = 38.255, p < 0.0001$). See Table C.4.

For instance, the synthetic Machiguenga agent responded "yes" 20% of the time to the question: *would you offer \$0 out of the amount endowed to the other player?* In contrast, the Aché agent accepted the \$0 offer only 3 out of 100 times ($p < 0.001$). To examine the differences in acceptance rates between each pair of groups, we performed Fisher's Exact Test and adjusted the p-values for multiple comparisons. See Table C.5.

Interestingly, our ChatGPT WEIRD benchmark rejected all zero offers across all iterations. While this finding contradicts the prevalence of zero modal offers observed in student populations [36], it aligns well with Mei et al.'s [10] findings that ChatGPT tends to avoid selfish behavior and exhibits greater generosity compared to human subjects in experimental settings.

The Yanomami tribe's offering pattern differs significantly from all other tribes except the Aché. Although both exhibit low acceptance rates for 0% offers, the Yanomami overall offer rates are low and closer to *homo economicus* compared to other tribes, with the Yanomami accepting offers only 4 times out of 1,100 trials.

B.2 Ultimatum Game

Previous research on economic behavior in small-scale societies has revealed important variations in ultimatum game (UG) outcomes. A study by Henrich et al. [25] found that the Machiguenga of Peru made low offers (mean 25%) with rare rejections (4.8% rejection rate), while the Orma of Kenya offered higher amounts (mean 31%) with lower rejections (4%). In contrast, the modal Aché offer was 50% and there were no rejections. The Tsimané showed no rejections as well, even for low offers, contrasting sharply with the Hadza who exhibited high rejection rates (24% mean, 43% modal for offers $\leq$ 20%)[2] (see Table C.2).

The differences of behavior in the UG are attributed to factors such as market integration, community size, and cultural norms of sharing and fairness [37]. The Machiguenga and Tsimané, for example, cooperate at the family level and it appears that the anonymity of the players in the UG removes fairness considerations. The Aché regularly share meat, which they distribute equally among all the households, irrespective of which hunter made the catch. In contrast, the Hadza's high rejection rates may be influenced by their tightly-knit communities, where fairness and equity norms are strongly enforced. In small-scale societies, low offers and high rejections are linked to a concept known as "tolerated theft" [39]. In these societies where resources are limited and sharing is expected, individuals may tend to make low offers to hold onto their resources, while also feeling entitled to a fair share of others' resources, leading to the rejection of low offers.

B.2.1 UG Proposer:

The acceptance counts out of 100 iterations for each offering level are shown in Table C.6. The data demonstrate that among the synthetic tribesmen, the Yanomami exhibit behavior most closely resembling *homo economicus*; all other synthetic tribesmen, including the Hadza, made generous offers. For most synthetic tribes and ChatGPT, the modal offers were 60% of the endowment, which is higher than the modal offers observed in experiments with human participants, as described in Table C.2. Despite of this, we found a significant association between group membership and the proposer's decision, accounting for the different offer levels (CMH $M^2 = 60.796, p < 0.001$).

B.2.2 UG Responder:

In contrast to observations with human tribesmen, synthetic tribesmen tend to have high rejections. However, the rejection rates show variability across tribes, largely following observed patterns of behavior in human subjects qualitatively.

We conducted a CMH test to examine the association between group membership and acceptance rates while controlling for offer levels (see Table C.6). The results suggest that the group a person belongs to (Ache, Orma, Tsimane, etc.) is significantly associated with their likelihood of accepting offers, even when taking into account the different offer levels (CMH $M^2 = 27.688, p < 0.001$). This could indicate cultural or other group-specific factors influencing decision-making in this context. Interestingly, the Yanomami exhibit high rejection rates, not consistent with *homo economicus* and ChatGPT rejects almost all offers that are $\leq 50\%$ of the endowment.

A chi-square test of independence, examining the relationship between tribal affiliation and the decision to reject low offers, indicates that rejection rates are significantly associated with tribal profiles ($\chi^2(5) = 36.389, p < 0.001$). Contingency tables for the chi-square tests are provided in Table C.8. For instance, the Hadza exhibit higher rejection rates than the horticulturalist Tsimané and Machiguenga (p $<$ 0.05). We performed Fisher Exact test to compare pairs of tribes. See Table C.9 for pairwise comparisons with adjusted p - values).

The results for UG responders are qualitatively consistent with observations with human subjects where Hadza are found to be more likely to reject low offers. The Yanomami, who as far as we know have not been subjected to experiments, show higher rejection rates than other tribes and almost as high as ChatGPT. Although the SCA representing the Yanomami offers little as dictator in the DG and as proposer in the UG, as responder, they reject a lot.

B.3 Endowment Effect

To our knowledge, no field experimental studies have investigated the endowment effect among the tribes other than the Hadza that we profiled in our experiments (see Table 4). This presents an opportunity to demonstrate that our methodology can be effectively used for piloting studies and refining experimental protocols with hard-to-reach populations.

Below we present the script of a conversation with an Aché bot upon taking a picture of a plate with two food items (temperature 0.65; max tokens = 150) – see Section 2 and Appendix A.2.3. This method of implementing the endowment effect is similar to condition 2 in Apicella et al's [26] experiment. We believe multimodal interfaces present a novel framework for experimenting with LLMs that more closely aligns with real settings. The Aché bot rejects the offer to exchange the endowed item.

```
Multimodal Endowment Effect

Hi!
Here are two food items. Do you recognize them?

I recognize them! They look like palm pith and fruit from the forest. We Ach\'e
    people love to gather and eat these foods in the forest. They are very
    nutritious and give us energy for our daily activities like hunting and
    foraging.

one of these fruits will be given to you

'Oh, that's very kind of you! I would be happy to receive either of the fruits.
    As an Ach\'e, I believe in sharing and generosity, so I would be grateful for
    whatever fruit you choose to give me. However, if I had to choose, I think I
    would prefer the palm pith. It's a staple food in our diet, and it reminds me
    of the forest and our traditional way of life.'

You are given the guava fruit
```

```
"Thank you so much for the guava fruit! I'm very grateful for your kindness. I'll
     make sure to share it with my family and friends when I return to our band.
    It's a delicious fruit, and I'm sure everyone will enjoy it. As an Ach\'e, I
    believe in sharing and generosity, so this gift will bring joy to many people.
     Thank you again for your thoughtfulness!"
```

```
that is ok, would you like to switch for the palm pith?
```

```
"No, that's okay. I'm happy with the guava fruit you gave me. As I said, I
    believe in sharing and generosity, and I'll make sure to share it with my
    family and friends. It's a kind gesture, and I appreciate it. Besides, I think
     it's good to appreciate what we have and not be too attached to specific
    things. The guava fruit will bring us joy, and that's what matters. But thank
    you for offering to switch!"
```

B.4 Rationale Examples

Examples of step-by-step explanations of decisions in the Ultimatum Game. The first entry is the proposer's "Yes" or "No" decision to offer \$6 to the recipient followed by an explanation. The endowment is \$10. The second entry is the recipient's "Yes" or "No" response to an offer of \$6 followed by an explanation. The complete set of responses can be accessed through our GitHub repository.

Synthetic Tsimané responses:

```
["No\n\n [EXP]\n\n1. As a member of the Tsimane tribe, I value cooperation,
    social identity, and community well-being.\n2. Offering $6 out of $10 would
    result in me keeping $4 and the other tribe member receiving $6,"],
["Yes\n\n [EXP]\n\n1. The offer of $6 out of $10 is more than an equal split, as
    half of $10 would be $5.\n2. By accepting the offer, I will receive a larger."
     portion of the money, which benefits me"],
```

Synthetic Hadza responses:

["No\n\n [EXP]\n\n1. As a member of the Hadza tribe, I would not offer $6 out of $10 in this experiment.\n2. The Hadza culture places a strong emphasis on equality and fairness, as evident in their communal sharing practices and egalitarian"],

["No\n\n [EXP]\n\n1.]. As a member of the Hadza tribe, we value egalitarianism and fairness in our social interactions.\n2. Accepting an offer of $6 out of $10 would mean I am receiving more than the proposer, which goes against"],

C Tables

Society	Mean Offer (%)	Median Offer (%)	Modal Offer (%)	Key Observations
Hadza (Tanzania)	20-30%	20%	0-10%	Lower offers in smaller camps, despite strong sharing ethic.[24]
Machiguenga (Peruvian Amazon)	n.d.	n.d.	n.d.	No specific data available on dictator game behavior.
Tsimane' (Bolivian Amazon)	41.2%	40%	50%	Significant variation based on village membership.[28]
Ache' (Paraguay)	n.d.	n.d.	n.d.	No specific data available on dictator game behavior.
Orma (Kenya)	40-50%	50%	50%	Offers influenced by market integration and community size.[30]
Yanomami (Amazon Rainforest)	n.d.	n.d.	n.d.	No specific data available on dictator game behavior.

Table C.1: Summary of Dictator Game Offers in Small-Scale Societies

Note: The table shows mean, median, and modal fractions of the endowment (usually a day's worth of work) offered to another member of the tribe. n.d. = no documented studies.

SS Society	Median (%)	Modal (%)	Mean Rej. Rate (%)	Mod. Rej. Rate (%)	Key Observations
Hadza (Tanzania)	26%	30%	24%	43% (for offers $\leq$ 20%)	High rejection rates, especially for low offers [**Henrich.2010**][25][47].
Machiguenga (Peruvian Amazon)	25%	15%	4.8%	10% (for offers $\leq$ 20%)	Very low rejection rates, even for low offers[27][25].
Tsimané (Bolivian Amazon)	30%	25%	0%	0%	No rejections observed, even for low offers.[**Henrich.2010**][25].
Orma (Kenya)	31%	30%	4%	0%	Low rejection rates influenced by market integration[**Henrich.2010**][25].
Aché (Paraguay)	40%	50%	0%	0%	No rejections observed, reflecting strong norms of sharing.[**Henrich.2010**][25]
Yanomami (Amazon Rainforest)	n.d.	n.d.	n.d.	n.d.	No specific data available on ultimatum game behavior.

Table C.2: Summary of Proposer Offers and Rejection Rates in the Ultimatum Game

Note: The table shows previous studies' proposer median and modal offers as a percentage of the endowment, as well as the mean and modal rejection rates for the recipient. n.d. = no documented studies.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Ache	3	2	1	0	2	0	1	1	2	1	0
Orma	11	3	4	0	5	4	1	2	3	1	1
Tsimane	21	3	2	5	1	3	0	7	0	2	2
Hadza	7	3	5	3	2	2	0	2	4	5	3
Machiguenga	20	4	4	1	3	4	1	2	2	3	0
Yanomami	1	0	0	1	2	0	0	0	0	0	0
Cochran-Mantel-Haenszel test results: $M^2 = 27.48, p - value = 1.586e^{-7}$											

Table C.3: Dictator's Agreement to Offer Rate (0% – 100%)

Note: The extremely low p-value indicates strong evidence against the null hypothesis that the common odds ratio across all strata is equal to 1. This suggests that there is a significant association between the cultural groups and the decision to accept/reject offers, after accounting for the various offer levels in the Dictator Game.

	Ache	Orma	Tsimane	Hadza	Machiguenga	Yanomami
Accept	3	11	21	7	20	1
Reject	97	89	79	93	80	99
Chi-square test results: X-squared = 38.255, df = 5, p-value = $3.354e^{-7}$						

Table C.4: Dictator's Decision to Accept or Reject an Offer Rate of 0%

Note: The p-value < 0.0001 suggests that the decision to accept or reject an offer rate of 0% is dependent on the tribe.

Comparison	p-value	Adjusted p-value	Significant ($\alpha = 0.05$)
Ache vs. Orma	0.0489	0.7335	No
Ache vs. Tsimane	0.0001	0.0015	Yes
Ache vs. Hadza	0.3311	0.8829	No
Ache vs. Machiguenga	0.0002	0.0030	Yes
Ache vs. Yanomami	0.6212	1.0000	No
Orma vs. Tsimane	0.0814	0.4070	No
Orma vs. Hadza	0.4595	0.9804	No
Orma vs. Machiguenga	0.1170	0.4388	No
Orma vs. Yanomami	0.0050	0.0175	Yes
Tsimane vs. Hadza	0.0072	0.0195	Yes
Tsimane vs. Machiguenga	1.0000	1.0000	No
Tsimane vs. Yanomami	0.000004	0.000060	Yes
Hadza vs. Machiguenga	0.0119	0.0317	Yes
Hadza vs. Yanomami	0.0649	0.3894	No
Machiguenga vs. Yanomami	0.000008	0.000120	Yes

Table C.5: Pairwise Comparisons of Acceptance Rates of 0% Offer in the Dictator Game

Note: We performed Fisher's Exact Test to examine the differences in acceptance rates between each pair of tribes. The results of the pairwise comparisons are shown under the p-values. The adjusted p-values represent correction for multiple comparisons using the Benjamini-Hochberg procedure for controlling the false discovery rate (FDR), which is the expected proportion of false positives among the rejected hypotheses.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Ache	1	4	15	28	18	29	51	47	21	12	1
Orma	0	4	4	23	26	42	63	56	37	9	0
Tsimane	1	2	4	17	26	30	38	38	14	10	1
Hadza	0	1	10	32	28	40	65	42	15	8	2
Machiguenga	0	11	7	28	41	56	75	60	32	12	8
Yanomami	0	5	5	13	11	15	15	0	8	0	1
Cochran-Mantel-Haenszel test results: $M^2 = 60.796, p - value = 6.328e^{-15}$											

Table C.6: Ultimatum Game Proposer Agreement to Offer Rate (0%-100%)

Note: The very small p-value we obtained from the Cochran-Mantel-Haenszel test provides evidence against the hull hypothesis of no association between tribe and proposer acceptance rates, controlling for offer rate. This indicates that group membership plays a crucial role in the decision-making process of the proposer in the Ultimatum Game.

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Ache	1	13	29	40	43	61	67	67	65	58	69
Orma	1	20	26	38	36	54	57	67	60	57	58
Tsimane	2	14	37	44	59	59	66	70	77	74	56
Hadza	2	13	14	43	52	83	54	69	68	65	83
Machiguenga	1	51	42	44	52	57	73	81	87	79	61
Yanomami	0	3	8	17	20	37	44	57	58	56	38
Cochran-Mantel-Haenszel test results: $M^2 = 27.688$, p-value = $1.426e^{-7}$											

Table C.7: Ultimatum Game Responder Agreement to Offer Rate (0%-100%)

Note: The very small p-value from the Cochran-Mantel-Haenszel test rejects the common odds ratio of acceptance versus rejection is the same for all groups across all strata (offer levels). Thus, group membership plays a crucial role in the responder's decision-making in the Ultimatum Game.

	Ache	Orma	Tsimane	Hadza	Machiguenga	Yanomami
Accept	27	28	32	23	46	9
Reject	73	72	68	77	54	91
Chi-square test results: X-squared = 36.389, df = 5, p-value = $7.941e^{-7}$						

Table C.8: Responder's Decision for Low Offers in the Ultimatum Game

Note: The p-value < 0.001 indicates that there is a statistically significant difference among the six synthetic tribes in their decision to accept or reject a low offer rate (10-30%). Values within the contingency table are averages of frequency counts corresponding to the three offer rates rounded to the nearest whole number.

Comparison	p-value	Adjusted p-value	Significant ($\alpha = 0.05$)
Ache vs. Orma	1.0	1.0	No
Ache vs. Tsimane	0.5353	0.6692	No
Ache vs. Hadza	0.6245	0.6896	No
Ache vs. Machiguenga	0.0080	0.0200	Yes
Ache vs. Yanomami	0.0015	0.0045	Yes
Orma vs. Tsimane	0.6436	0.6896	No
Orma vs. Hadza	0.5166	0.6692	No
Orma vs. Machiguenga	0.0125	0.0235	Yes
Orma vs. Yanomami	0.0009	0.0037	Yes
Tsimane vs. Hadza	0.2050	0.3075	No
Tsimane vs. Machiguenga	0.0592	0.0986	No
Tsimane vs. Yanomami	0.0001	0.0006	Yes
Hadza vs. Machiguenga	0.0010	0.0037	Yes
Hadza vs. Yanomami	0.0113	0.0235	Yes
Machiguenga vs. Yanomami	4.07e-09	6.11e-08	Yes

Table C.9: Pairwise Comparison of Low Offer Acceptance Rates in the Ultimatum Game

Note: We performed Fisher's Exact Test to examine the differences in acceptance rates of low offers (10% - 30%) between each pair of tribes. The results of the pairwise comparisons are shown under the p-values. The adjusted p-values represent correction for multiple comparisons using the Benjamini-Hochberg procedure for controlling the false discovery rate (FDR), which is the expected proportion of false positives among the rejected hypotheses.

D RAG vs Fine-Tuning in Cultural Profile Generation

In our study, we opted for a Retrieval-Augmented Generation (RAG) approach rather than fine-tuning language models for each cultural group. This decision was based on both theoretical considerations and practical implications for our methodology. Here, we discuss the differences between RAG and fine-tuning in the context of generating cultural profiles for Synthetic Cultural Agents (SCAs).

D.1 Theoretical Considerations

RAG leverages the powerful in-context learning capabilities of large language models (LLMs). This allows the model to adapt to new information provided in the prompt without modifying its underlying parameters. In our case, this means we can create cultural profiles for various tribes by simply providing relevant information in the context, rather than training separate models for each culture.

While fine-tuning has been shown to improve model steerability, allowing for more precise control over the model's outputs, our primary concern is with recall - the ability to accurately retrieve and utilize specific cultural information. RAG excels in this aspect by directly incorporating relevant information into the generation process.

Moreover, RAG potentially allows for better generalization across different cultural contexts. By keeping the base model unchanged and modifying only the retrieval corpus, we maintain the model's broad knowledge while focusing on specific cultural details.

D.2 Practical Implications

Our RAG approach offers greater flexibility in studying various cultural groups. We can easily update or modify the information used for different tribes without the need to retrain entire models. This is particularly advantageous when working with evolving cultural information or when expanding to new cultural groups.

Fine-tuning separate models for each tribe would require significant computational resources and large corpora of text for each cultural group. Given the limited textual data available for many small-scale societies, this approach would be impractical. RAG allows us to make efficient use of the available information without the need for extensive training data.

By using RAG, we maintain a clear separation between the general knowledge encoded in the

LLM and the specific cultural information we provide. This allows for greater transparency in our methodology and gives us more control over the exact information used to generate cultural profiles.

Furthermore, our RAG methodology is more accessible to other researchers who may want to replicate or extend our work. It doesn't require specialized model training and can be implemented using publicly available LLMs and retrieval techniques.

D.3 Relevance to Our Methodology

In our study, the use of RAG aligns well with our goal of creating versatile SCAs that can represent a wide range of cultural groups. By injecting cultural information into the context rather than the model parameters, we maintain the flexibility to rapidly prototype and refine our cultural profiles.

This approach also allows us to clearly trace the sources of cultural information used in generating SCA behaviors, enhancing the interpretability of our results. Furthermore, it enables us to easily update cultural profiles as new information becomes available or as cultures evolve over time, without the need for retraining.

While fine-tuning could potentially offer more specialized models for each cultural group, the trade-offs in terms of data requirements, computational resources, and reduced flexibility make RAG a more suitable choice for our current methodology. As the field progresses and more extensive cultural datasets become available, future research could explore the potential benefits of fine-tuned models in comparison to our RAG approach.