Designing a Multiplayer Competitive Text-Based Game for Elicitation of Tacit Knowledge About Humor from Crowds of People

Melvin Schop[1] , **Agathe Balayn**[1] , **Ujwal Gadiraju**[1] , **Jie Yang**[1]
[1]Delft University of Technology

Abstract

Tacit knowledge, unlike explicit knowledge, is not easily codifiable, yet important for machine learning models. This research explores a method to gather tacit knowledge about humor using a simple text-based party game, building on the existing idea of using games to gather tacit knowledge from crowds of people. Players propose prompts, which will then be answered by other players. They will then vote to determine which of the two answers to each prompt is the funniest. The engagement of the players with the game is measured and tacit knowledge is obtained from the jokes. With a large and diverse enough group of participants across games, a variety of tacit knowledge can be extracted.

1 Introduction

Machine learning is becoming increasingly popular and important in today's world. Applications that involve object recognition, artificial synthesis of plausible scenes, or other similar fields often require tacit knowledge to perform tasks with higher accuracy. Explicit knowledge can be easily collected and often consists of simple facts that can be retrieved from widely available databases. Tacit knowledge on the other hand is often more cultural, intuitive, or otherwise not easily expressible in simple facts. Where humans have this tacit knowledge at their disposal from birth and through experience in the real world, artificial intelligence may need to be taught or provided such knowledge to perform more accurately. An example of a task that is easy for humans because of tacit knowledge, but harder for machines would be trying to distinguish serious statements from joking statements, or understanding the context within the real world from just a couple of sentences. Ritchie (2009) [6] goes into great detail as to what has been possible by computers through classical methods. With the use of tacit knowledge, these kinds of tasks could be improved significantly. This knowledge may be gathered from people at a large scale using gamification methods. The focus of this paper will therefore be on various gamification methods to improve the collection of tacit knowledge at a large scale.

Work by Von Ahn and Dabbish (2008) coins the term Games with a Purpose (GWAPs) [1]. These games explore gamification techniques and game design strategies to reach a certain goal efficiently and accurately. Von Ahn et al. (2006) [11] demonstrates a way to extract common-sense knowledge using a single-player game. These previous works will form a starting point for this paper, further specializing the idea in a competitive multiplayer setting using text-based games, namely, whether a certain answer to a prompt could be considered funny or not. It is also important to quantify the knowledge that is collected in some way. Yatskar et al. (2016) [12] describes various methods to quantify common-sense knowledge using a word-net. The type of tacit knowledge that this paper will focus on is tacit knowledge about humor using simple prompt-answer jokes. Sjöbergh and Araki (2007) [8] use word space models to determine humor based on closeness to other words often found in joke texts. Sjöbergh and Araki do this without trying to understand the underlying meaning of the text. This may be an option to classify differences between different bubbles of people. These bubbles can be thought of as groups of people that can be characterized by some common factors, such as region, culture, or age. However, keeping this tacit knowledge raw will allow machines to learn directly from examples and in turn, become more capable of generating or detecting various types of jokes similar to the jokes present in the collected data. The generated jokes could be tailored to different bubbles of people for use in chatbots for psychological treatment or better recognition of jokes in other automated chats to generate more relevant replies. The ability for machines to have a better understanding of jokes may make the experience of interacting with machines feel more akin to human-to-human interaction. Collecting tacit knowledge about humor will be an interesting case to research further, which may also open up paths to collecting tacit knowledge about other topics.

The focus of this research is to explore a different method for collecting large amounts of qualitative tacit knowledge with a focus on engagement. This method could be adapted for the collection of various types of tacit knowledge. Namely, we intend to answer the question, if a text-based multiplayer competitive game is sufficiently effective at acquiring tacit knowledge about humor from crowds of people for use by machine learning models. We intend to answer this question through these sub-questions:

- Is the game sufficiently engaging for the involved players?

- What tacit knowledge can we extract from free-text answers?
- How reliable is this method?
- How does this compare to other methods?

We try to solve this problem by designing an engaging multiplayer game to collect tacit knowledge about humor. The experience of the game as well as the quality of the data are important. The idea of using humor may form a simple basis to increase engagement and aid in the overall quality and quantity of data. By not requiring any external databases to be used, this game design may allow for many different types of tacit data to be collected with different topics and different metrics.

The methodology describes some sub-problems regarding the collection and processing of the data. The game design will be further explained and the metrics used to describe how and what quantitative and qualitative data is captured. The game flow and the experimental setup are precisely stated with the requirements of each phase regarding data collection and their function within the game. A short section about responsible research will describe some of the ethical considerations about data collection and the assurance of the data quality and reproducibility of the experiment. The discussion will focus on comparing this method to previous works and describe some of the limitations of the chosen method. The final section will go over the conclusions and any potential future work.

2 Methodology

Collecting data for machine learning models consists of three sub-problems: The first problem is to collect data, the second problem is to anonymize the data, and the third problem is to analyze and store the data in a way that is useful for machine learning models. The focus of this paper will be on the first problem, as the second and third problems already have various existing methods that try to solve these problems and are therefore out of the scope of this paper.

Before we can collect tacit knowledge about humor, we need to have an idea of what humor is. Veatch (1998) [10] goes into great detail about the theory of humor, describing various common aspects within humor and various types of humor that are recurring throughout many different cultures. It is hard for machines to have a true understanding of humor, namely, what makes something funny and under what circumstances. The main idea to help collect tacit knowledge to improve this understanding is to design a simple prompts-and-answers game for small groups of people. The game will allow the players to produce answers to prompts related to various topics. The goal is for the players to produce answers that they think could be considered funny by other players. Since pairs of answers are contested, each player should aim for the most humorous answer they can think of to obtain a higher score. With these prompts, answers, and votes, we could gather knowledge about what could be considered funny surrounding various topics, which answers to prompts are best and further create potential to distinguish aspects about jokes between bubbles of people.

The design of the game will be very simple and allow for implementations shaped as web-based applications, standalone apps that connect through the internet, simple games that can be played in the real world using props or even be embedded in existing games such as Minecraft, Jackbox Games or other platforms that include a form of multiplayer and/or creative input. The data will be gathered from all players and processed externally using a simple script.

The type of data collected could be personally identifiable. The data could be anonymized through various methods, but there are some potential issues to consider that could make anonymization difficult or potentially introduce biases in the data or alter the data in unwanted ways:

- **Wording**: The vocabulary used could identify players.
- **Typos**: Spelling mistakes, typos, or capitalization could be identifiable.
- **Identity Literals**: The data could include names of people as part of the punchline.

Some of these issues could be avoided by removing any capitalization and punctuation, auto-correcting typos or spelling mistakes, anonymizing names, etc. These manipulations to the data, however, could alter the punchline of jokes in unwanted ways. Thomson et al. (2005) [9] describes why this is the case. This type of data is sensitive to any alterations and fully anonymizing the data will be very difficult. Anonymization techniques for microdata as described by Ghinita et al. (2009) [4] are difficult to apply to qualitative data. The data is hard to quantify, due to the free-text format. Simple actions such as turning the data into lowercase or removing punctuation could result in unwanted changes in the meaning of the text. The question "What do you call a crossbreed between an elephant and an ant?" with the answers "An elephANT" and "an ELEPHANT" for example would be combined into a single data point after removing any capitalization. However, as both answers carry a different message, with the first answer putting a clear emphasis on the "ant" part, this would be incorrect. Since the scope of this research does not extend to processing natural language, the players will be informed in advance that all their prompts and answers will be stored as-is.

The game is implemented in Minecraft, as this is accessible to a lot of players across a diverse player base. This is not an entirely new idea, as others like Singh (2020) [7] have been using Minecraft as a platform to research and experiment on for AI purposes in the past. Duncan (2019) [3] also describes Minecraft as more than just a game. Adding to "exciting experiments in games for learning" (Duncan, 2019), Minecraft can be a great platform for getting people together to play a game for research purposes, while the general look and feel of Minecraft makes it feel casual and helps players feel like they are playing against each other through a virtual avatar, rather than a web app which could make players feel less in touch with other players. The implementation will also be quicker, as handling connections, sessions and various other low-level problems are already part of Minecraft's infrastructure and the code can be focused purely on the game logic. Players can easily be kept track of and the game's infrastructure further allows for an easy collection of data and statistics. Some

technical limitations could cause roadblocks in the implementation. However, the aforementioned benefits outweigh these limitations.

3 Prompts and Answers

The main goal of this experiment is to test the effectiveness of a specific type of game to collect large amounts of tacit knowledge for use with machine learning models. The main idea is to let groups of people compete against each other in a party game that lasts up to 10 minutes per round at most.

Experimental work

In this game, each of the players will have to come up with prompts that could lead to funny answers, but these prompts may also be provided by the game itself if players fail to submit them. The players will then continue to answer these questions and finally vote for their favorites. This design allows the players to evaluate the answers of other players within the closed ecosystem of the game. Engagement will be measured through the active participation of players in the game. A database of prompts, answers, and votes will also be compiled. To categorize or label this data at a large scale, Natural Language Processing (NLP) methods should be used. The work of Jing et al. (2018) [5] could be an option, or equally the work of Kevin et al. (2012) [2]. Their NLP solutions may work sufficiently well on the data that is generated in this experiment. However, for the small amount of data that is collected during this research, setting up and validating the NLP systems would be more time-consuming than manually analyzing the data. Choosing not to use automated systems to clean up the data also further keeps the data from being altered and any findings based on the data can be directly tied back to the context of a game session.

Quantitative Data

Engagement can simply be measured by keeping track of the time or the number of rounds played. However, a broader concept of engagement will also include how much fun the player had, if they will play the game again later, or if they will recommend the game to others. Engagement is all about interaction. The more engaged a player is with the game, the more that player will interact with the game. In the case of this game, the types of engagement will mostly be captured through an after-game survey. Additionally, the amount of games played and spectated is kept track of. It is furthermore possible to see the number of votes that have been cast or received per player. The rounds played can be a quantitative metric while the surveys serve more as a qualitative metric. The combination of these metrics could be used to express how engaging the game is.

Qualitative Data

Answers can be ranked by the amount of positive interaction they have received compared to other answers to infer a ranking of answers per prompt. The amount of votes is dependent on the number of total votes that could have been received for that answer. Because combining similar answers is another non-trivial NLP problem in the context of jokes as explained earlier, it is important to keep track of how many votes the answer could have had to determine a ratio of votes received to the total votes it could have received. If the game is played with only 3 players, for example, the maximum amount of votes that could be received is 1 vote per round, while with 10 players, there could be a maximum of 8 votes per round. The only downside is that the granularity and weights of these ratios are not well preserved. Any errors or ambiguity in the votes is much more apparent for smaller amounts of votes, a score of 8/8 will be much more accurate than a score of 1/1. It is therefore important to keep the tally counts separate and include a maximum vote count per round to preserve this information.

4 Experimental Setup and Results

The experiment consists of three parts: The setup of the game and which parameters were used for time limits. The extraction of the data from the game, which includes any cleaning and formatting of the data. And lastly, the results with an analysis of the quantitative and qualitative data.

4.1 Game Setup

The game is designed to work for 3 up to 10 players that may participate directly and any amount of spectators who only have the power to vote. Players may communicate through voice or in-game chat if they wish. The game consists of several phases: Lobby, Prompt, Answer, Vote and Results. The time limits may be shorter if all players submit their responses early. Character constraints include a character limit that is either fixed by character width or by character count, but allows any printable ASCII characters and most Unicode characters to be part of the input. This allows for non-Latin scripts to be used as well. The players are identified by an index from 0 to N-1, where N is the amount of participating players. This identifier will also be used by the prompts and answers that they generate or the substitutes thereof in case of players dropping out or failing to submit within the time limit. The votes will still be counted towards the score of the player responsible for the input even if their input was substituted. Additionally, people can join the lobby and enter as spectators or queue up for a game with more than 10 players. From the pool of queued players, 10 random players will be selected who will actively participate in the game, the rest of the players will be spectating for that round. Spectators are also able to vote for answers in the voting phase like regular players. Each phase is situated in a different room and players are separated during the prompts and answer phases. These phases are described in more detail below.[1]

Lobby

The *Lobby* phase allows players to connect, the rules can be explained and players can get ready before the game commences. When the game is initiated, all players that participate will be assigned a random ID from 0 to N-1 and tagged as playing. Players will not know which ID has been assigned to them, nor any of the other players' IDs, as these IDs are completely internal. If they leave the game at any moment during

[1] see Appendix A3 for screenshots of the game

this session, they may reconnect and resume at any moment to be returned to the game at the proper current phase. The offline players' actions will be substituted with sensible data. What this sensible data entails is further described per phase.

Prompt

The *Prompt* phase allows players to insert a prompt, which can be an open question or a statement containing blanks that players will be able to respond to with a variety of answers. All players are shown the same topic, which they should use to base their prompt upon. Among these topics are: Work, Food, Animals, People, etc. The full list of topics can be found in Appendix A2. The text field theoretically allows players to insert up to 798 characters, based on the width of the thinnest character. However, in practice, using the average width of letters in human languages, a player could insert around 270 characters effectively. Players are given roughly 120 seconds to come up with a prompt and submit it. Prompts that have not been submitted or empty prompts will be discarded. A discarded prompt will be substituted by a randomly picked prompt selected from a database compiled from previously collected prompts. A simple game setting allows for this phase to be skipped, which effectively forces the previously collected prompts to be reused, such that the pool of answers for these prompts could be further saturated.

Answer

The *Answer* phase rotates the prompts from the previous phase and distributes them among the players. This process is deterministic through the following algorithm:
Let the successor be k+1 for any k from 0 to N-1 mod N, which makes the identifier 0 be the successor to the identifier N-1. The prompt for any player i from 0 to N-1 will go to the successor (i+1) and their successor ((i+1)+1). This will work for any integer N greater than 2.
Effectively making sure that each player is served two prompts from two distinct players without getting their own prompt. They are given 120 seconds to answer both of these prompts with an answer. The character limit for a single answer is 50 characters. Answers that have not been submitted within the time limit will be discarded and replaced by an empty string. It is possible for the first answer to be submitted exclusively without the second answer being submitted. The player will always answer the prompts in the same order starting with their predecessor's prompt.

Vote

The *Vote* phase allows players to vote for the answer that is the funniest answer to a prompt to their best beliefs. The prompts will be displayed to all players and spectators. The answers to that prompt are displayed to either side and can be clicked on to vote for that particular answer. Players cannot vote in case either of the two answers to that prompt was created by them. Each player can vote at most one time for either one of the answers per prompt. The maximum amount of votes that can thus be given is N-2 votes by players not including additional votes from spectators. A player can choose to abstain from their vote by letting the timer of 30 seconds (per prompt) run out or by clicking the skip option. Spectators are also able to vote, but they will have to submit their votes before the active players do. This is to keep the pace of the game up in the event there are many spectators.

Result

The *Result* phase will show a leader board with the winning player(s) on a pedestal. This phase lasts 30 seconds, during which a random distribution of fireworks is displayed. After this phase, everyone is returned to the lobby.

4.2 Data Extraction Setup

The relevant data that should be extracted includes a list of anonymized player statistics. This includes the number of votes received and cast, and the number of games played and spectated. Furthermore, the data should be extracted per game, with a list of prompts and a maximum amount of votes that could have been given per prompt. Per prompt, the two answers with the number of votes they received should be stored. The data is not further processed as described earlier.

4.3 Results

The data collected has a large qualitative part and a smaller quantitative part. The qualitative part mostly consists of raw player inputs such as prompts and answers to prompts. The quantitative part is obtained from measuring usage statistics as well as a short after-game survey filled in by the participating players.

Qualitative Analysis

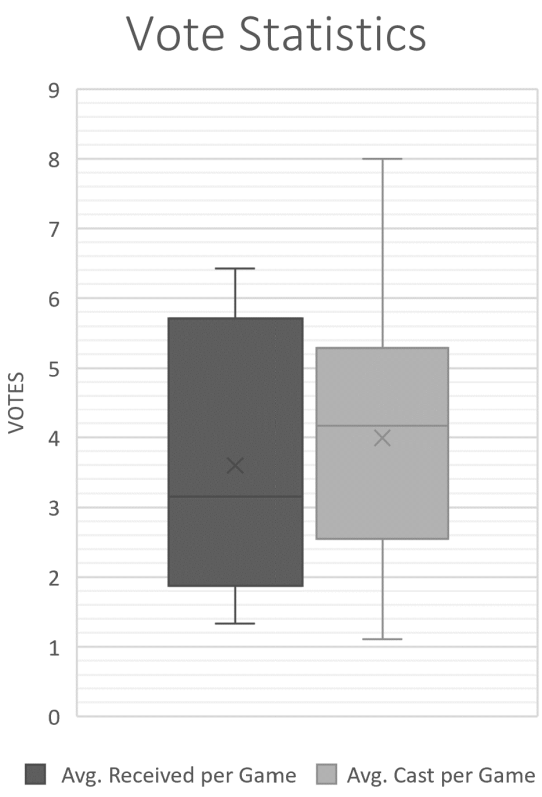

Figure 1: Distribution of vote averages among players

A total of 15 different players contributed to the raw data consisting of 121 distinct prompts and 371 total answers to these prompts collected throughout 40 games. The total of votes that have been cast/received is 582. The average vote distribution per player can be seen in Figure 1. Of all the votes that could have been cast, only 8% were skipped, which means the total of votes that could have been cast would be around 630 votes with an error margin of about 2 votes. The reasons why players decide to skip votes seem to range from "the answers are too similar" to "neither of the answers is funny". This could mean that the prompt was faulty, difficult to understand, or does not inspire players to produce creative answers. This is a perfect example of such a prompt:

> **Q**: "What's the best sex position?"

One pair of answers to this prompt yielded:

> **A1**: "69, for obvious reasons" *(1/3 votes)*
> **A2**: "" *(0/3 votes)*

The second option being blank results in the first answer obtaining a single vote.
Another pair of answers to this prompt yielded:

> **A1**: "69, obviously..." *(0/3 votes)*
> **A2**: "69 ;)" *(0/3 votes)*

Which has all 3 possible votes skipped. Any other answer would more likely have yielded a higher amount of votes. An example of an interesting prompt would be:

> **Q**: "What would be the name of the first country on the moon?"

This resulted in 3 pairs of answers:

> **A1**: "Amoonica" *(3/3 votes)*
> **A2**: "a" *(0/3 votes)*

> **A1**: "Zimbabwe 2" *(0/3 votes)*
> **A2**: "LUNAR LAND(ing)" *(3/3 votes)*

> **A1**: "Something from Greek mythology I guess" *(2/3 votes)*
> **A2**: "moonanistanland" *(1/3 votes)*

There is a clear pattern for the first two sets of answers leaning towards the answer related to the subject of the prompt, namely, the moon. For the third set, the answer that ties into some tacit knowledge surrounding the subject of the moon gets more votes instead of the answer that contains the word "moon". Because many objects in space are assigned various unique names, this could be a play on the Roman names of gods for planets such as Venus, Mars, etc. This may also just be pure coincidence, however.

Some of the data suggests the votes are hard to quantify. For the prompt:

> **Q**: "What do you call a crossbreed between a crocodile and a flamingo?"

These sets of answers were given:

> **A1**: "A croccodingo" *(0/1 votes)*
> **A2**: "Pink death machine" *(1/1 votes)*

> **A1**: "Crocodingo?" *(3/3 votes)*
> **A2**: "Pink Fluffy Alligator, Dancing on Rainbows" *(0/3 votes)*

But as can be observed, the very similar first answer either got all the votes or none of the votes. While the second answer contained "Pink" both times, it is very difficult to quantify any of this data with only 2 sets of answers. Potentially with 10 or more sets, you could start to see patterns in the voting. It is not enough data to draw any general conclusions.

Engagement
Players who filled in the engagement survey mostly reported positively[2]. The character limits were reportedly a bit too tight for some people.

-	What parts of the game were most engaging?
voting	3
answering	3
"everything"	2
unanswered	2
-	What parts of the game were least engaging?
waiting	3
creating prompts	2
voting	1
unanswered	4

Figure 2: Engagement feedback from all 10 participants who filled in the survey

Figure 2 shows that the most engaging parts of the game include voting and answering for the most part, while less engaging parts of the game include the waiting and the creation of prompts. Among the suggestions for the game, only two were about the actual game design. One comment that mentioned an even quicker pace of the game would be easy to satisfy by simply reducing the time limits. However, this would also make it more difficult for players to come up with prompts and answers in time, as in some cases these answers were not submitted. Most players thus say that the time limit is just right. Other suggestions include better decorations and quality of life fixes for problems that arise due to the limitations of the nature of Minecraft's architecture.

5 Responsible Research

Gathering data from players in games can be a controversial topic. It is important that the privacy of the people involved is highly valued and any usage of their data is justified for research purposes. All players that participated in this research have been asked to fill in a short form that asks if they consent to the data being used in this research. Any data that has been made public is anonymized or processed first such that no data points can be traced back to their origin. The players were only informed about the rules of the game and that

[2] see Appendix A1 for the complete survey results

their data will be used for research purposes. The exact goal of the research was not emphasized. This is done to reduce bias due to a potential change in behavior such as overthinking or fitting their inputs to match the intents of the research and instead allow for more genuine inputs.

The concept of humor is not static and may change rapidly over time or could be perceived differently by different bubbles of people. Therefore, it is extremely difficult to precisely reproduce results derived from these human inputs, especially at a small scale. Most of the people involved in this experiment were part of a bubble with similar characteristics. Since the amount of data is relatively small and does not include randomly selected players, the conclusions drawn from the qualitative data are very limited and cannot be generalized. It is possible to do a case-by-case analysis of the data points to come to new insights which may stimulate further research. Reproducing the experiment with similar groups of people carefully selected by a host would be a viable option to gather results that will be of a similar form. It is possible to gather some general knowledge from the entire set of data to come to some basic conclusions as to which answers may be funnier than other answers within the scope of this particular bubble of players. The quantitative data surrounding engagement will mostly be reproducible, however. Although, due to the nature of this project, the smaller size of the group of participants and the lower diversity thereof should still be taken into consideration.

An implementation in Minecraft could introduce hidden variables or biases, namely, it would limit participants to a specific bubble of people. Limitations within the game could also introduce a bias. The code is made public[3] to ensure a reproducible environment for this platform. This allows for an easy way to reproduce the environment and reduce any bias that could have resulted from implementation differences. Further technical instructions specific to that implementation are provided on the repository.

6 Discussion

Unlike the work of Yatskar et al. (2016) [12], the qualitative data collected in this research is less refined than theirs. One could consider using a method described by Sjöbergh and Araki (2007) [8] to further extract keywords after applying NLP methods to reduce the amount of data. This could result in further refined quantitative metrics. However, with the relatively small amount of data collected, conclusions from such quantitative metrics would be less meaningful as the rate of errors and biases would include too much noise to produce conclusive results. The collected knowledge may still be used to provide samples for machine learning models to create new relations between prompts about certain topics and all possible answers to these prompts can be supplied with weights on quality and tags for context. By providing a raw set of data, however, the complete game sessions are captured. This may result in more accurately trained models with large quantities of knowledge. Filtering the data could potentially leave out the details, which could be important in the aforementioned scenarios with regards to the scope of humor.

[3] source code: https://github.com/AgentM12/funny-not-funny

Comparing the results about engagement with Ahn et al. (2006) [11], the average time played per person is significantly higher, although the amount of players involved is much lower, due to various reasons. One of the reasons is that the game was privately hosted. A public host and distribution of the game would have been possible given more time and more channels to advertise through. Ahn et al. state an average of 29 facts per player, while our game states around 71 facts per player on average. It should be noted, however, that the type of data collected in our game is different from theirs. Despite this, the similarities between our game and theirs are still high enough to make a comparison of engagement relevant.

Implementing this game in Minecraft has many implications on what is possible to collect in terms of data and how the game is presented to the players. Minecraft is a popular platform for AI research as the work of Singh (2020) [7] has inspired us to attempt to apply Minecraft as a platform for research to a different field within computer science. Without the use of any 3rd party tools, running a plain Minecraft server has its benefits, but also its limitations. On the one hand, gathering data is rather easy, as the infrastructure is already there in the form of scoreboards and statistics. The game logic can also be hot-fixed relatively easily while the game is running. On the other hand, it does require a proficient understanding of the internal coding language of Minecraft. Since Minecraft works with commands rather than pure code, each command is rather verbose and takes up an entire line of code. Logic and data management is only secondary. Due to this nature, the style of programming may be more cumbersome than conventional programming languages. Manipulating data like strings is not supported in Minecraft. This does not pose a big problem, however, since the data can be easily exported and processed externally.

Party games are often experienced as a fun gathering with lots of laughing and jokes. By letting people compete in such a casual environment, the data they will produce will be genuine and simple to understand and not contain too many forced jokes or jokes that take a lot of time to understand. These jokes will more likely be thought of on the fly. These types of jokes often fall in a similar category of jokes which will make the data for machine learning more specific. The format of jokes could be further restricted to allow for an even tighter set of data, which in turn may specialize the underlying machine learning models further.

Most of the conclusions were drawn from the quantitative data. Some of the conclusions are drawn from the qualitative data in conjunction with the quantitative data in a more general sense. Even though the examples given are representative samples from the data set, the data is too sparse and lacks diversity for a proper qualitative analysis. Any of the findings with specific data points are merely hypotheses that could be tested in further research but do not have a proper scientific basis to truly draw conclusions from for this paper. The data of individual players has been sufficiently anonymized and raw statistics have been measured. These statistics are simple facts that can then be interpreted to come to some conclusions.

The results could be biased due to the nature of this project.

With limited reach and time, the groups of people involved in this experiment are either direct or indirect friends or players from a similar discourse. Deployment of the game in a public place would be a possible solution to reach a higher diversity and larger amount of players. The experiment could be conducted with strangers by advertising the game through various channels. The game could be scaled up to work in scenarios with larger amounts of players to help diversify the types of people in the experiment as well as collecting larger amounts of data. With a larger size and diversity in data, there would be a more accurate and scientific basis to conclude from, namely with regards to the generality of jokes in terms of location and time and potential clustering of jokes on these metrics using already known methods.

Another possible concern with players that know each other is that some players tend to vote for their favorites based on their relationship with the player. It is a valid concern that has been apparent from some data points that had to be removed from the database for this precise reason. These mention names of players directly or particular animals that they favor. This could make machine learning models believe that "cats" are inherently a funny answer, however, this could also come from favoritism by players. Therefore conducting this experiment with strangers further may aid in the collection of lesser biased data from which more concrete conclusions could be drawn. However, these biases are interesting when we look at differences between groups of friends playing this game, which could in turn deliver more topics to be researched in the future.

The data for engagement seems to agree with the hypothesis that using a platform such as Minecraft with a group of players can be very enjoyable. The added benefits of having a virtual room with players' avatars make players more willing to wait. Players will also feel more connected through the in-game chat or voice chat, which further helps with keeping players engaged in the game. Without any communication between players, this sense of competition with other players is lost and the feeling that one is simply inputting data into a machine seems more likely.

7 Conclusions and Future Work

Conclusions

We intended to answer our main question: "How effective is a text-based multiplayer competitive game at acquiring tacit knowledge about humor from crowds of people for use by machine learning models?" by looking at sub-questions such as: "Is the game sufficiently engaging for the involved players?", "What tacit knowledge can we extract from free-text answers?", "How reliable is this method?" and "How does this compare to other methods?". The conclusions to these questions are described below in further detail.

The level of engagement is high among the participants measured according to their behavior producing inputs, their playtime, and the survey testing for engagement. The amount of data gathered for the size of the experiment is reasonable. An average of 6 participating players per 10 rounds will yield around 120 answers within about an hour. This could be easily scaled up to run in parallel by using different hosts to allow for a large and diverse set of data to be collected in a relatively short period. The game is easy to understand which also helps to improve the quality of the data and the engagement of the game. Interestingly enough, a lot of the jokes will only be understood among all or some players within that bubble of players. Players tend to tailor their answers more to what they believe will be perceived as funny by (some of) the others within that group. This could create a possibility for comparison between two different groups on the premise of in-bubble humor. Having fun also contributes to the engagement of the game. A bubble of closely related players is more likely to play for extended periods as well as returning to the game periodically.

The types of tacit knowledge that can be collected are numerous. Relations between prompts and jokes or relations between bubbles of people and their style of humor may be collected. Collecting large amounts of qualitative data to be used in machine learning models is feasible. However, the data is still raw and techniques to refine this data may diminish the overall quality of the data.

The experiment may produce varying results depending on many variables. Due to the nature of free-text input, the constraints on the format of data that is entered by players can not easily be regulated. This can be easily solved by making the prompts fixed by carefully constructing prompts and propose these to the players, which will in turn also reduce the overall length of a game.

Future Work

We demonstrated that data collected through this type of game is relatively easy and overall engaging to players. Only a small subset of potential questions surrounding this subject has been tested. Improvements can be made in the selection of players, the platform, various tweaks in how the game flow is presented, and any potential changes in restrictions for the players, such as different time or character limits, or even a specialized category of topics as a common ground for the players to base their prompts on. Gathering a larger quantity of data may allow for more specialized conclusions such as determining various contexts in which jokes could be considered funny. This may require personally identifiable data to be used such as region, gender, religion, political beliefs, etc. to be able to say anything about differences in humor between bubbles of people. The idea of hosting the game through Minecraft could be adapted to other games which are inherently engaging. Other games may be accessible to different groups of people and may yield different results. Furthermore, the implementation could be improved by using third-party tools, such as plugins, to allow for more efficient data collection, as well as implementing methods to share prompts and answers between different server hosts in a common database. Third-party tools further allow customizability of the game experience within Minecraft and could aid in stricter text constraints.

8 Acknowledgements

We would like to thank Nadine Gruber for proofreading this paper many more times than we ever could have asked for.

Special thanks for the numerous reviews and critical feedback on this paper by peers of this project with similar topics within this field: Mădălin Broscăreanu, Andy Hu, Neha Kalia, and Jeffrey Lim.

References

[1] Luis von Ahn and Laura Dabbish. Communications of the acm. *Designing Games With a Purpose*, 51(8):57–67, 2008.

[2] Kevin Crowston, Eileen E. Allen, and Robert Heckman. Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, 15(6):523–543, 2012.

[3] Sean C. Duncan. Minecraft, beyond construction and survival, Oct 2019.

[4] Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Trans. Database Syst.*, 34(2), July 2009.

[5] Xiaonan Jing, Chinmay Talekar, and Julia Taylor Rayz. Comparing jokes with nlp: How far can joke vectors take us? In Norbert Streitz and Shin'ichi Konomi, editors, *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts*, pages 310–326, Cham, 2018. Springer International Publishing.

[6] Graeme Ritchie. Can computers create humor? *AI Magazine*, 30(3):71, Jul. 2009.

[7] Sameer Singh. Minecraft as a platform for project-based learning in ai. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13504–13505, Apr. 2020.

[8] Jonas Sjöbergh and Kenji Araki. Recognizing humor without recognizing meaning. In Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, editors, *Applications of Fuzzy Sets Theory*, pages 469–476, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[9] Denise Thomson, Lana Bzdel, Karen Golden-Biddle, Trish Reay, and Carole A Estabrooks. Central questions of anonymization: A case study of secondary use of qualitative data. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 6, 2005.

[10] Thomas C Veatch. A theory of humor, 1998.

[11] Luis von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 75–78, New York, NY, USA, 2006. Association for Computing Machinery.

[12] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, San Diego, California, June 2016. Association for Computational Linguistics.

A Appendix

A.1 Tables and figures

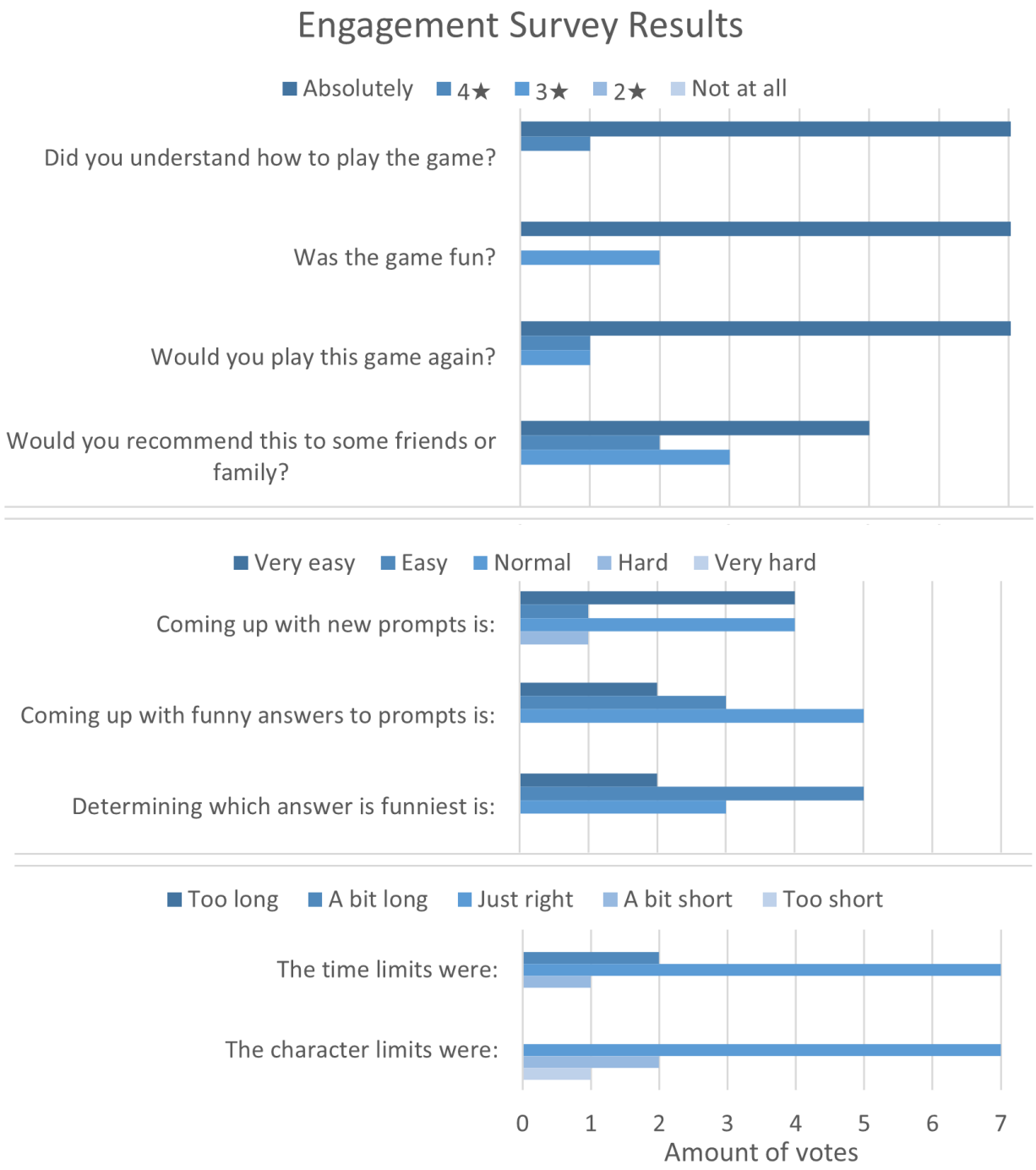

Figure 3: Survey results of 10 participants

A.2 Data

Topics

The complete list of topics used in the experiment. This is shown to the players in the prompts phase of the game and could be used by the players to base their answers upon:

> Work, School, Traffic, Transportation, Food, Movies, Holidays, Internet, Aliens, Space, Ocean, Shopping, Language, Fishing, Love, Medical, Weather, Technology, Sleep, Nature, Kids, TV Shows, Fantasy, Minecraft, Games, Clothes, Animals, People, Countries, Music, Family, Friends.

A.3 Screenshots

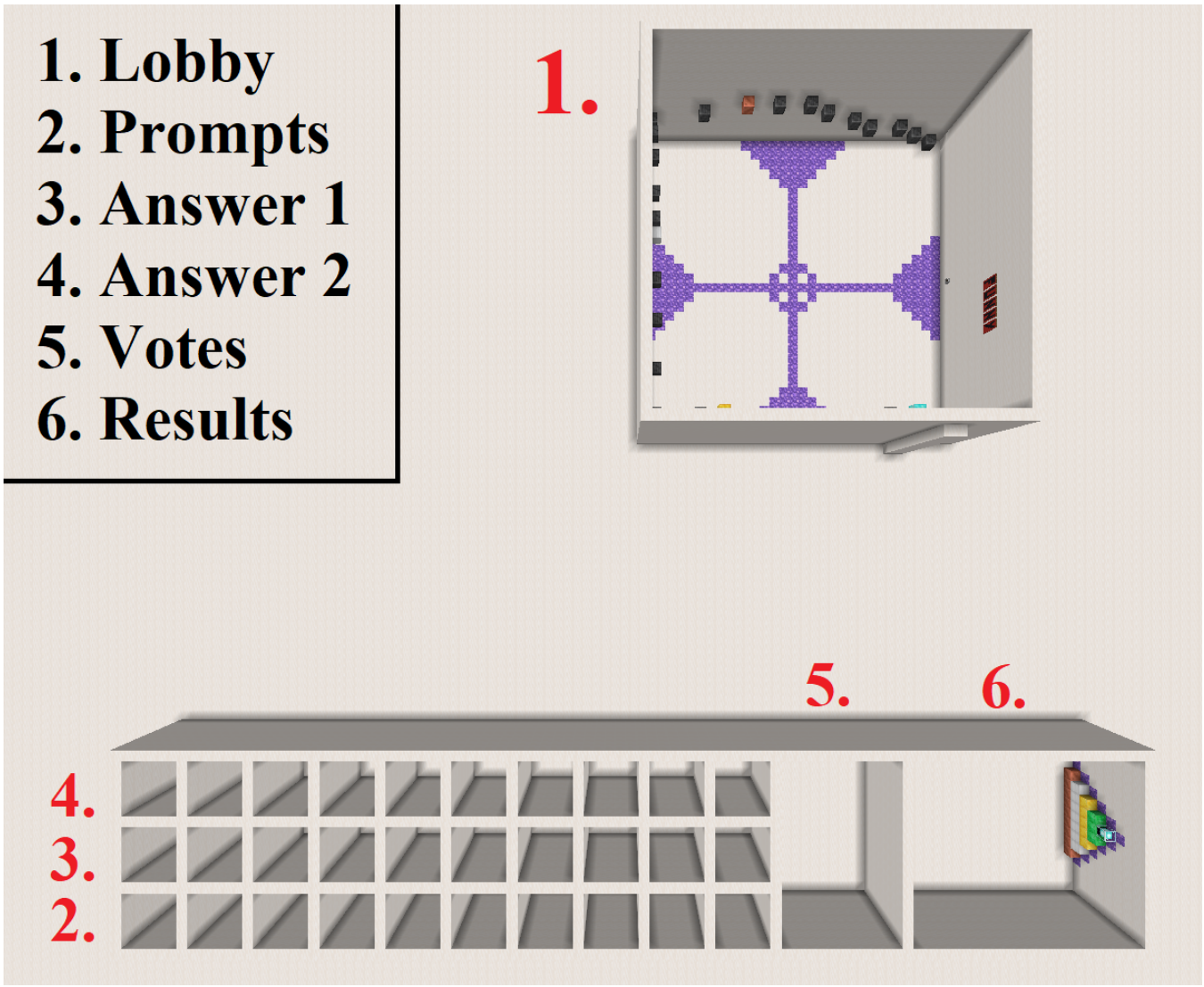

Figure 4: The phases that facilitate the general flow of the game

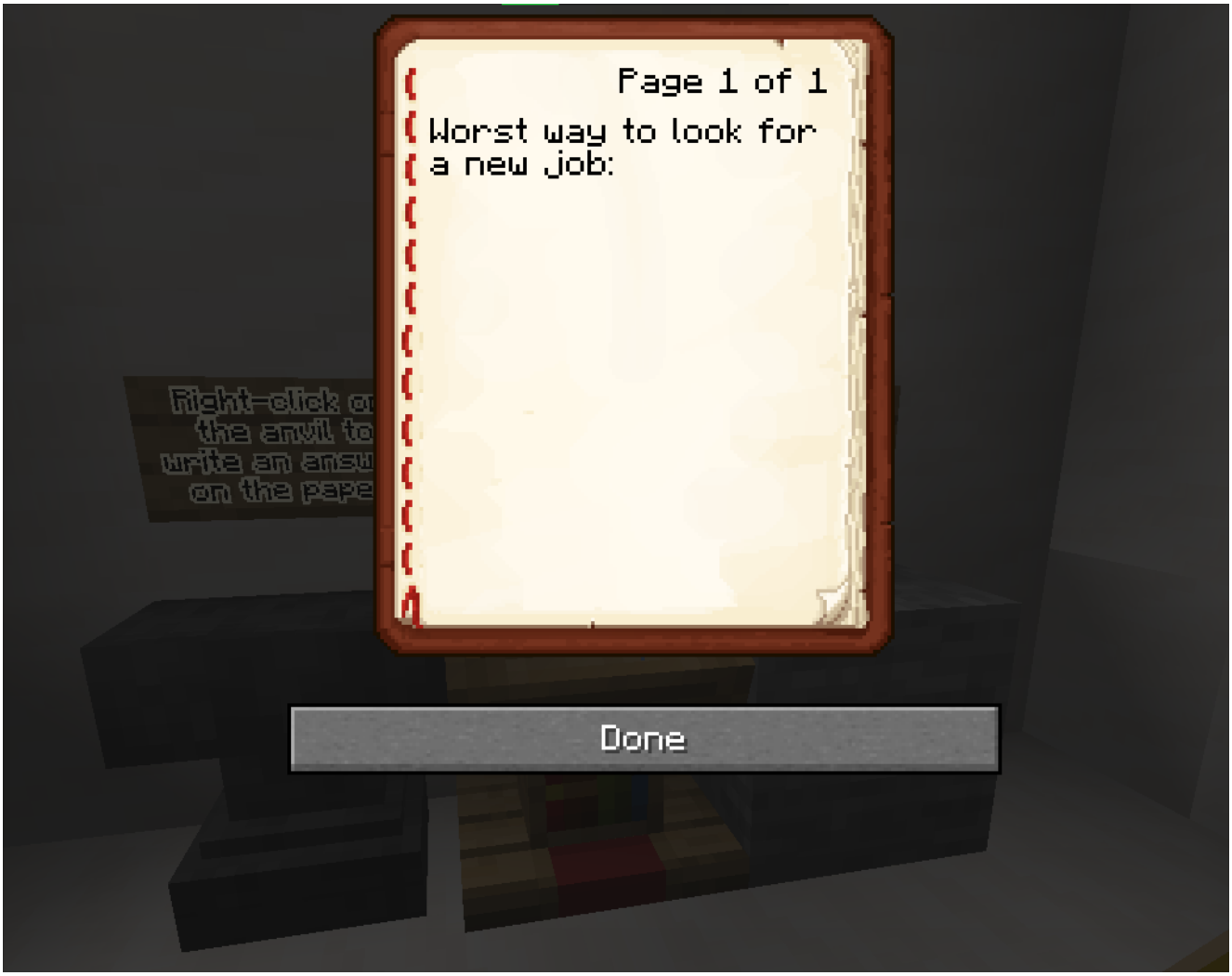

Figure 5: Reviewing a prompt can be done by accessing a lectern

Figure 6: One can vote by interacting with signs

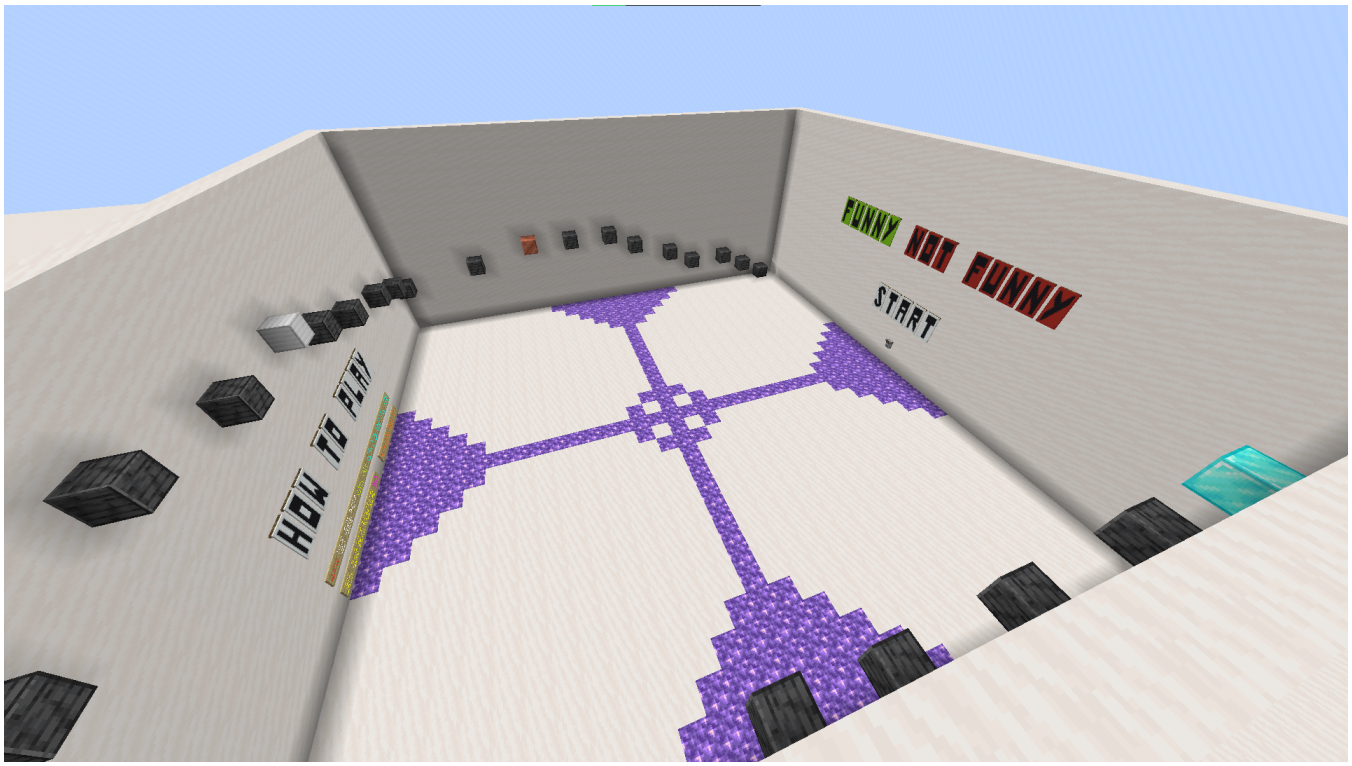

Figure 7: A simple yet engaging lobby with a parkour to ease waiting and a quick how-to-play guide.