# Exploring Differences Between Hybrid GPT-Human and Human-Created Qualitative Codebooks in an Educational Game

Xiner Liu[1[0009-0004-3796-2251]], Zhanlan Wei[1[0009-0002-3931-6398]], Amanda Barany[1[0000-0003-2239-2271]], Jaclyn Ocumpaugh[1[0000-0002-9667-8523]], Ryan S. Baker[1[0000-0002-3051-3232]], Andres Felipe Zambrano[1[0000-0003-0692-1209]], Yiqiu Zhou[2[0009-0000-9004-9307]], and Camille Giordano[1[0009-0001-2772-3079]]

[1] University of Pennsylvania, Philadelphia, PA 19104, USA
xiner@upenn.edu
[2] University of Illinois, Champaign, IL 61820, USA

**Abstract.** This study explores the ability of GPT-4 working together with humans to generate a codebook to analyze scientific observations from middle school learners in the What-if Hypothetical Implementations in Minecraft (WHIMC) project. It compares this Hybrid codebook to one fully developed by Humans using a variety of techniques to evaluate how the codes developed by each approach relate to one another and to external measures of student interest. Results show that the Hybrid GPT-Human codes consist of broader categories that align more consistently with the external interest metrics, whereas the Human codes offer finer-grained insights into specific student behaviors. However, the complementary insights offered by each suggest that combining both approaches could improve our understanding of student engagement and inform more effective strategies in educational game design and intervention.

**Keywords:** Large Language Models, GPT, Automated Codebook Development, Codebook Comparison, Epistemic Network Analysis

## 1 Introduction

A key early step in most Quantitative Ethnography (QE) analyses is qualitative coding, which involves defining key concepts and identifying where they occur [28] in a variety of data types (e.g., interview data, [39]; game logs, [14]; open-ended surveys, [17]; and online discussion posts, [12]). Coding procedures facilitate systematic data organization, allowing researchers to understand, compare, and replicate patterns of complex phenomena like learning and motivation (e.g., [29], [30]). This systemization is particularly important in game-based learning contexts, where the design of the learning experiences can facilitate complex and nonlinear patterns of engagement (e.g., [31]).

As datasets grow in scale and artificial intelligence (AI) grows in capacity, researchers have increasingly explored how automated tools such as large language models (LLMs) can enhance qualitative codebook development (e.g., [38]) and apply codes to

data (e.g., [33]). LLMs have complementary strengths to other approaches used in the QE community, such as nCoder+ (e.g., [37]).

Hybrid codebook development processes involving LLMs appear to yield a wider variety of high-quality codes than fully-human processes [1]. Questions remain, however, about how the content of codes differs between hybrid and fully-human codebooks, and how this influences the eventual research outputs. In response, this work offers an exploratory evaluation of the insights that a Hybrid codebook—developed by humans supervising an LLM's development process—might yield within a complex game-based learning context. We explore these questions within the context of data on students' scientific observations within WHIMC (What-If Hypothetical Implementations in Minecraft; [15]), an educational game that leverages Minecraft to engage students by immersing them in interactive, hypothetical astronomy scenarios (e.g., *What if Earth had no moon?*). WHIMC combines playful learning with scientific inquiry, encouraging learners to collect data on habitability factors, generate hypotheses, and compare conditions across virtual worlds.

Specifically, we compare and contrast two distinct codebooks developed to understand the scientific observations that students are making within WHIMC. The first was fully developed by human researchers, and the second was created with a hybrid approach where GPT-4 provided the initial codes, which were then discussed and refined by the research team. We then examine the affordances of each approach, by comparing the constructs across the two codebooks and inter-rater reliability (IRR) within each codebook, followed by a correlation analysis to measure their relationship with game-external measures of student interest. Lastly, we applied Epistemic Network Analysis (ENA) to each codebook to compare the findings yielded by each approach.

## 2 Related Work

Traditionally, codebooks are developed manually by human researchers, who draw on domain expertise to identify and refine categories [26]. This process, while effective, is labor-intensive and may be more difficult for researchers to conduct reliably at scale [8]. For years, others have sought to automate this process using natural language processing methods (e.g., [25]), but this has often resulted in codes that were limited in quality and complexity [10]. In addition, prior tools could not explain their reasoning [22]. Large Language Models (LLMs) such as ChatGPT offer improvements by leveraging advanced natural language processing to interpret and generate text in a human-like way from open-ended inputs, supporting a dialogue between researcher and tools that can enhance, challenge, and supplement human understanding (e.g., [13]).

Recent work has explored possible best practices for hybrid human-AI codebook development. For example, [7] use GPT-3.5 to extract and group codes from qualitative data, which are refined through discussions with a human coder. The GPT-generated codes aligned closely with those identified by the human coder. Similarly, [23] leverage LLaMA-2-70B-Instruct to identify themes and generate initial codes from interview data, which are subsequently adjusted through multiple iterations with domain experts. The LLM-generated themes show moderate to substantial alignment with human-

generated themes. Barany et al.'s work assessed the quality of codes produced using four different methods: fully manual, fully automated, and two different hybrid approaches combining human input with ChatGPT [1]. Their findings showed that a hybrid approach where ChatGPT refined the initial codebook developed by humans resulted in more reliable, higher-quality codes compared both to fully automated and fully manual methods.

These advances matter for research in game-based learning because understanding student experience often requires more than quantitative measures alone. As students navigate these immersive worlds, they generate a rich array of data, which includes not only quantitative behavioral metrics such as in-game actions [34] and sensor data [6], but also qualitative data such as in-game chat messages [4] and written responses [11]. While quantitative methods are effective for tracking performance patterns [32] and behavioral trends [16], they often fail to capture the full nuance in data such as students' conversations [24] and reflection [5]. Qualitative analysis is particularly valuable for interpreting the narrative, reflective, and collaborative aspects of gameplay [21], which show not only what students do but the cognition and perceptions that accompany their behavior. Central to this type of analysis is the use of codebooks, which reflect researcher perspectives on the important categories within the data, and guide how researchers label the data. Because codebooks serve as the backbone of qualitative data analysis, comparing codebooks developed using different methods for analyzing the same in-game data reveals how various approaches capture the complex, qualitative dimensions of student engagement, which helps us better interpret students' interactions and meaning-making within the game.

## 3 Methodology and Results

### 3.1 Data

This study analyzes data from 32 middle-schoolers who used WHIMC during summer camps held at three schools across different U.S. regions, representing urban, suburban, and rural areas. Of these, 19 are male, 11 are female, and 2 either identified as a third gender or did not disclose their gender. The group included 10 White, 14 African American, 2 Native American or Pacific Islander, and 3 students from other racial or ethnic backgrounds. Participation was voluntary, with both students and their parents providing written consent before the camp began.

The data collection process spanned five days and included several surveys designed to assess student interest. On day one, students completed an Boeder et al's science interest development survey [3]. On days one through three, they explored Minecraft environments (e.g., Earth, Earth with no moon) designed to simulate different planetary conditions and introduce them to astronomy concepts. On day three, students completed two surveys developed by Gadbury [9], for astronomy interest and Minecraft interest. On day four, the interest development survey was re-administered to assess any changes in students' interest levels and Linnenbrink-Garcia et al.'s situational interest scale was administered [18]. In total, 22 students completed all four surveys, which used structured questions and rating scales to capture students' interest levels at different stages.

During the camps, students explored scientific concepts, which were incorporated into the WHIMC curriculum to encourage active learning about the habitability of virtual planets WHIMC simulates. With in-game tools, students measured environmental factors like temperature, humidity, and gravity, and recorded their observations (e.g., *"The gravity is 0.1657 times the Earth's gravity."*) on wooden signs. These observations were visible to other learners within the game and served as the basis for our exploration of Human versus Hybrid codebook development.

### 3.2 Fully Human Codebook Development

Our fully Human codebook development builds upon prior work that sought to better understand student observations in WHIMC [19, 36]. Specifically, we examine a codebook that was developed to better understand the degree to which student observations reflected WHIMC's educational goals [36] but was expanded upon with an inductive thematic analysis [19], to better explore student interest development. Previous study has applied this codebook to analyze observations made by learners within the game and found that learners who maintained or increased their astronomy interest often produced more complex observations, such as making comparisons and forming hypotheses, compared to those whose interest declined [19]. In the current study, we consider six of these codes, excluding two codes that rarely occurred in the data (<5%). The finalized version of the developed codebook is presented in Table 1.

**Table 1.** Abbreviated Codebook Developed by Human Researcher

| Code Name | Definition |
|---|---|
| Measure/ Descriptive | Def: Observations related to measures of physical attributes of different planets and moons, including temperature, quantity, weight, etc. Ex: *"the pressure is 14.48 psi"* |
| Comparison | Def: Observations that compare or contrast conditions among in-game worlds or between real-life experiences and in-game worlds, including instances of violated expectations. Ex: *"the pressure is lower than Kepler 186"* |
| Hypothesis | Def: Observations that make hypotheses or guesses, showing speculative thinking, conjectures, or predictions. Ex: *"solar panels to generate energy for heat, light, etc."* |
| Questioning | Def: Observations involving meaningful questions about game mechanics or world elements or reflecting curiosity to understand the game. Ex: *"do the plants thrive in this lighting?"* |
| Noun | Observations stating nouns without any elaboration (previously labeled as "factual" in [36]. Ex: *"there is a windmill"* |
| Exclamations | Observations of pure exclamations without accompanying explanations. Ex: *"i found birds!!!!"* |
| Social | Observations of interactions with other players through communication or shared exploration. Ex: *"[Redacted] doesn't need oxygen"* |
| Non-game | Observations of sequences of characters, emojis, or symbols repeated excessively, random numbers or letters, or unrelated references to movies, books, and celebrities. Ex: *"vfbfgbtgnt"* |

### 3.3 Hybrid Codebook Development

Two additional researchers, not involved in the development of the initial human-developed codebook, used ChatGPT-4 to create a separate, Hybrid codebook using the same dataset as described above. Specifically, we first used ChatGPT to generate an initial list of constructs and definitions, which was then refined by human researchers for alignment with the research goals. Automating the early stages of codebook development has several advantages [1]: it reduces overall development time, leads to higher reliability between human coders when applied, and results in codes that receive higher quality ratings from independent, code-blinded human evaluators than either fully manual or fully automated approaches [1]. Following Barany et al.'s framework for automatic codebook development [1], this process began with an initial data review to understand the context and structure of the WHIMC dataset. Then, human researchers provided ChatGPT with a subset of the data to prompt the generation of constructs. After multiple rounds of prompt engineering, a prompt was identified that produced consistent coding results, with no more than two constructs differing across repeated tests on a 100-line sample set:

> *Hi ChatGPT, you are a great qualitative researcher. I want you to analyze the following comments and observations made by students while playing an educational version of Minecraft with different worlds focusing on astronomy. Please develop a qualitative codebook by conducting a thematic analysis to identify the main themes that emerge from the "observations" across different worlds pertaining to students' science learning in the game environment. The structure of the codebook should include names, definitions, example codes, and explanations for the codes. I will divide the data into chunks of 100 lines and give you one chunk at a time. You should use all chunks of 100 lines to do the thematic analysis. With every chunk, you can update and edit the codebook if needed to make sure it still captures all the themes across the previous chunks. Here is the data:*

This prompt was then applied to the entire dataset, which was divided into 18 batches of 100 lines each to comply with the model's token limits. ChatGPT generated around eight themes per batch, but only themes that appeared at least twice across batches were retained. Overlapping themes were consolidated into broader categories through researcher discussions with the clearest, most comprehensive definitions selected or merged. No new codes were proposed by humans during this process. The resultant codebook is outlined in Table 2. A previous study by [35] applied this codebook to examine how students' levels of situational interest relate to the content of observation notes they posted during gameplay. The study found that students with higher interest recorded more observations and more often referenced science-related concepts than those with lower interest.

**Table 2.** Abbreviated Codebook Developed by ChatGPT with Human Refinement

| Code | Definition |
|---|---|
| Astrophysical and Planetary Observation (*Astrophysical*) | Observations of planetary bodies' physical characteristics, including visual and environmental details. Ex: *"this planetary object does not rotate"* |
| Scientific Inquiry and Reasoning (*Scientific*) | Observations where logical reasoning, hypothesis formation, questioning, or understanding of scientific concepts. Ex: *"is there a 24 hours without a moon"* |
| Socioemotional Responses (*Socioemotional*) | Observations reflecting emotional states and social reactions to game events. Ex: *"IM ON TOP THE WORLD!!!"* |
| Environmental Adaptation and Survival Strategies (*Environmental*) | Observations of strategies and actions developed to survive and adapt to environmental conditions. Ex: *"lifes would be shorter* a lot *of heat strokes will happen in my opion humans cant live here"* |
| Resource Identification and Utilization (*Resource*) | Observations of interactions with and identification of game resources, including discussions on their use. Ex: *"this must be a different sleeping area"* |
| Cultural and Playful References (*Cultural*) | Observations of mentions of popular culture or playful elements indicating engagement beyond educational goals. Ex: *"IT IS DECEMBER!!!!!!!!!!!!!"* |

### 3.4 Interrater Reliability

Both the Human and Hybrid-developed codebooks were independently applied by human raters (the first and third authors for the Human-created codebook and the fourth and eighth authors for the Hybrid codebook), to a 200-line training dataset. Cohen's Kappa ($\kappa$) was used to assess IRR, and any discrepancies were resolved through social moderation. Reliable human agreement ($\kappa$>0.70) was achieved for all constructs prior to social moderation, across both the Human codebook (avg $\kappa$=0.85) and the Hybrid codebook (avg $\kappa$=0.78). Observation data was then automatically labeled using GPT-4o, accessed through OpenAI's API, using prompts proposed by [20]. For both codebooks, automated coding was implemented using either Zero-Shot Prompting, in which GPT-4o was provided with construct definitions, and Few-Shot Prompting, in which GPT-4o was provided with both definitions and a small number of labeled examples to illustrate how each construct should be applied. Both approaches were evaluated against the human-coded ground truth, and the prompting approach that achieved the highest performance for each coding category was selected for use in the final coding process. The resulting IRR between GPT and human labels was comparable to that between the two human raters ($\kappa$>0.70, avg $\kappa$=0.83).

### 3.5 Conceptual Overlap Across Codebooks

After applying the two codebooks, we assessed the level of conceptual overlap between constructs from the two codebooks using Cohen's Kappa (note that although Kappa is often used for IRR, this is not the purpose that Kappa was used for here), by seeing the degree to which each data point received each code. This enabled us to identify which constructs share thematic similarities across the two codebooks. Although developed independently, Figure 1 shows that the two codebooks converge on several core themes.

For example, the high overlap between *Questioning* and *Scientific* ($\kappa$=0.65) suggests that both codes capture the idea of observations where learners are actively seeking to understand or investigate, despite the specific terminology used. Similarly, the high overlap between *Exclamation* and *Socioemotional* ($\kappa$=0.57) reflects a shared identification of expressive, emotional responses as a distinct theme in learner interactions. Lastly, the relatively strong overlap ($\kappa$=0.45) between *Non-game* and *Cultural* suggests that both codes capture external or culturally influenced references that extend beyond the core learning tasks.

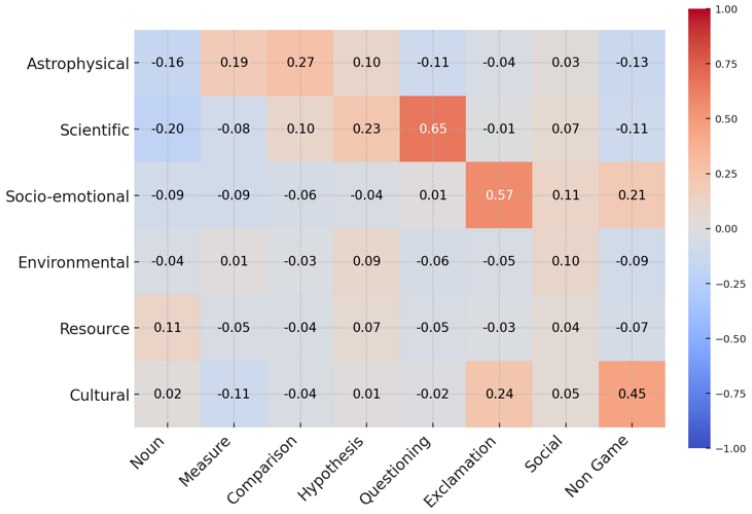



**Fig. 1.** Heatmap of Kappa Scores Between Constructs in The Two Codebooks

### 3.6 Correlation between Constructs and Survey Measures

We next compared how each codebook's constructs correlate to interest survey measures. This analysis provides insight into how different patterns of in-game observations, represented by the constructs in both codebooks, are associated with the students' interest development throughout the camp. After coding the data using both codebooks, we summed the occurrences of each construct at the student level and calculated their correlation with survey-measured interest using Spearman's method. We selected Spearman's correlation because the survey data did not follow a normal distribution, and this method is more appropriate for capturing monotonic relationships.

Our study explores the correlations between four distinct interest survey measures and the constructs in two separate codebooks, leading to 24 statistical tests for the Hybrid codebook (4 surveys × 6 constructs) and 32 for the Human codebook (4 surveys × 8 constructs). Due to the large number of tests run, we used Monte Carlo analysis to determine whether the overall number of significant results was more likely than chance and then used the Benjamini-Hochberg (B&H; [2]) procedure to control the false discovery rate (FDR) for individual tests. The Monte Carlo analysis conducted 10,000

simulations for each codebook and randomly generated binomial outcomes based on the number of tests and the chosen significance level of 0.05. For each simulation, we compared the number of significant results to the observed values for both codebooks (5 for the Hybrid codebook and 1 for the Human codebook). We then calculated the p-values for each codebook by determining the proportion of simulations where the number of significant results met or exceeded the observed number of significant results. We then compared the p-values obtained from the statistical tests against the adjusted alpha thresholds for each codebook, based on the 24 and 32 tests, respectively, for B&H.

Prior to conducting B&H correction, there were only 2 constructs out of 32 that were statistically significant within the Human-created codebook. The Monte Carlo analysis yielded a p-value of 0.49, suggesting that the overall pattern could have occurred due to chance. The relationship closest to significant in the Human-created codebook (see Table 3) was the correlation between *Questioning* and Astronomy Interest (r=0.50), meaning that students with more developed astronomy interest ask more questions.

In contrast, the Hybrid codebook yielded 5 statistically significant results (out of 24). According to the Monte Carlo analysis, the probability of observing five or more significant results out of 24 tests by random chance is approximately 0.005. Three Hybrid constructs showed significant correlations with situational interest even after B&H correction: *Scientific* (r=0.54), and *Resource* (r=0.62), while *Astrophysical* (r=0.48) was only marginally significant after the B&H correction. In other words, students who report higher situational interest engage more frequently with scientific concepts, explore more extensively with planetary characteristics, and generally pay more attention to in-game resources. Two other Hybrid constructs were correlated with Astronomy Interest: *Scientific* (r=0.57) and *Resource* (r=0.50), meaning that astronomy interest is associated with both noticing and reasoning with the scientific elements in the game.

**Table 3.** Spearman Correlations between Construct Occurrences and Interest Measures

| Code | Source | Interest Dev. Gain | Situational Interest | Astronomy Interest | Minecraft Interest |
|---|---|---|---|---|---|
| **Resource** | Hybrid | 0.16 (p=0.48) | 0.62 (p<0.01)* | 0.50 (p=0.02) † | 0.40 (p=0.06) |
| **Scientific** | Hybrid | 0.05 (p=0.82) | 0.54 (p<0.01)* | 0.57 (p<0.01)* | 0.30 (p=0.17) |
| **Astrophysical** | Hybrid | 0.03 (p=0.89) | 0.48 (p=0.02)† | 0.32 (p=0.15) | 0.25 (p=0.27) |
| **Socioemotional** | Hybrid | -0.18 (p=0.43) | 0.28 (p=0.20) | 0.07 (p=0.77) | 0.08 (p=0.71) |
| **Cultural** | Hybrid | -0.09 (p=0.68) | 0.19 (p=0.39) | 0.03 (p=0.89) | 0.02 (p=0.93) |
| **Environmental** | Hybrid | -0.23 (p=0.31) | -0.03 (p=0.91) | -0.26 (p=0.24) | -0.12 (p=0.59) |
| **Social** | Human | -0.2 (p=0.34) | 0.48 (p=0.20) | 0.53 (p=0.10) | 0.36 (p=0.10) |
| **Comparison** | Human | -0.03 (p=0.91) | 0.45 (p=0.04) | 0.17 (p=0.45) | 0.16 (p=0.47) |
| **Questioning** | Human | 0.01 (p=0.96) | 0.41 (p=0.06) | 0.50 (p=0.02) | 0.33 (p=0.14) |
| **Measure** | Human | 0.04 (p=0.87) | 0.36 (p=0.10) | 0.18 (p=0.43) | 0.20 (p=0.36) |
| **Noun** | Human | 0.24 (p=0.29) | 0.33 (p=0.14) | 0.40 (p=0.06) | 0.38 (p=0.08) |
| **Hypothesis** | Human | -0.27 (p=0.22) | 0.28 (p=0.21) | 0.08 (p=0.74) | 0.05 (p=0.82) |
| **Exclamation** | Human | -0.07 (p=0.77) | 0.27 (p=0.23) | 0.23 (p=0.29) | 0.17 (p=0.45) |

* Significant *after* B&H correction
† Marginally significant *after* B&H correction

Finally, we compared the proportions of positive correlations between the two codebooks by applying Stouffer's Z test, which combines individual p-values from each codebook into a single Z-score. Specifically, Stouffer's Z can be used to compare the difference in Z-scores based on p-values by subtracting Z values, dividing the result by the square root of 2 (the number of tests being compared), and converting back to a two-tailed p-value [27]. Our analysis shows that the Hybrid codebook produced marginally significantly more positive relationships than the Human-created codebook, $Z=1.81$, $p=0.07$. The marginally greater proportion of statistical significance found for the Hybrid codebook suggests that the constructs identified by this codebook may capture patterns in student observations that align more closely with the development of interest than those identified by the Human-created codebook, where the associations appear weaker and more diffusely related to student interest measures.

### 3.7 Insights from ENA

Building on the insights we gained from the correlation analysis, we next examined the two codebooks using Epistemic Network Analysis (ENA), a method allows for the visualization and comparison of relationships between constructs within the same dataset. Mapping the co-occurrence patterns of codes using ENA offers insights into how different constructs connect and whether the application of each codebook influences our understanding of learner processes. This method can uncover subtle conceptual distinctions or overlaps that may not be evident from frequency counts or correlation matrices alone. To construct the networks, the data was grouped (unit variables) and segmented (conversation variables) by student and class day to capture connections during continuous periods of play. A moving window length of 4 was used to associate codes in each line of student data to their prior 3 lines; this window was chosen based on an interpretive review of how far back each student tends to refer to prior information in their observation discourse.

Two separate epistemic networks were created for extreme cases—one for the player with the lowest interest (Tiger, pseudonym) and one for the player with the highest interest (Cat, pseudonym)—to show how each codebook produces different interpretations of the same student. Tiger reported the lowest gain in Interest Development (-2.5), the lowest Situational Interest (0.94/7), and the lowest Minecraft Interest (0.5/5). Therefore, we chose to focus on this student for further analysis to better understand how each codebook might reveal or omit specific constructs that could potentially be associated with lower engagement and interest in this particular student. In contrast, Cat reported the highest level of interest across all individual interest surveys: Situational Interest (4), Minecraft Interest (4), Astronomy Interest (5/5), Interest Development (5.9→6; though the net increase was not the highest due to consistently high scores).

Figure 2 shows the two epistemic network models for Tiger, based on Human and Hybrid codebooks, reveal distinct patterns of engagement during gameplay in WHIMC. The ENA using the Human-created codebook shows a strong connection between *Non-*

*game* and *Noun* (line weight, lw=0.80), suggesting that Tiger's interaction with the game involved frequent categorization, potentially lacking the depth necessary to sustain meaningful scientific learning. Additionally, the prominent presence of non-game elements, observed 35% of the time throughout Tiger's observations, indicates a possible off-task behavior or distraction, where the student's attention drifted away from core astronomy-related goals, which is consistent with his low scores across all interest measures. Notably, two of the higher-order scientific activities (*Comparison* and *Hypothesis*) are absent from Tiger's model.

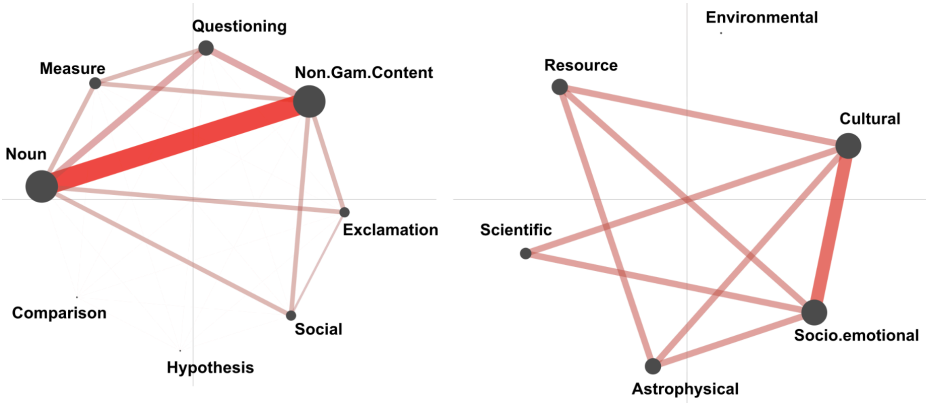



**Fig. 2.** Tiger's Epistemic Network for Human-created Codebook (*right*) and Hybrid GPT-Human Codebook (*left*)

Similarly, in the epistemic network produced with the Hybrid codebook, the most prominent connection involves codes that are not scientific in nature (*Socio-emotional* and *Cultural,* lw=0.60). Additionally, seven of the eight connections involve *Socio-emotional* or *Cultural* nodes, which reflects that while Tiger was emotionally involved (e.g., 35% of Tiger's observations included single words written in uppercase, such as *"WATER,"* which often reflect amplified emotional expression), this engagement may have been diffuse rather than focused on core scientific inquiry. This suggests that Tiger's engagement within the game was more centered on playful or emotional elements rather than sustained, curiosity-driven learning about astronomy.

Figure 3 shows the epistemic networks for Cat, whose higher interest is reflected in these visualizations. In the Human codebook's model, the stronger connections involving *Questioning* (e.g., *Comparison,* lw=0.73; *Hypothesis*, lw=0.46) suggest that Cat is intentionally engaging in sense-making and inquiry (e.g., *"why is the tides so low?"*). The Hybrid codebook's model also identifies this pattern, where strong connections associated with *Scientific* (e.g., *Astrophysical,* lw=0.88; *Resource*, lw=0.39) highlight that Cat is forming hypotheses and asking probing questions (e.g., "*it is warmer underground because they have airflow?"*). Though each codebook labels these behaviors differently, both highlight Cat's active efforts to interpret and understand the game environment, driven by curiosity and a desire to explore the game's content and mechanics.

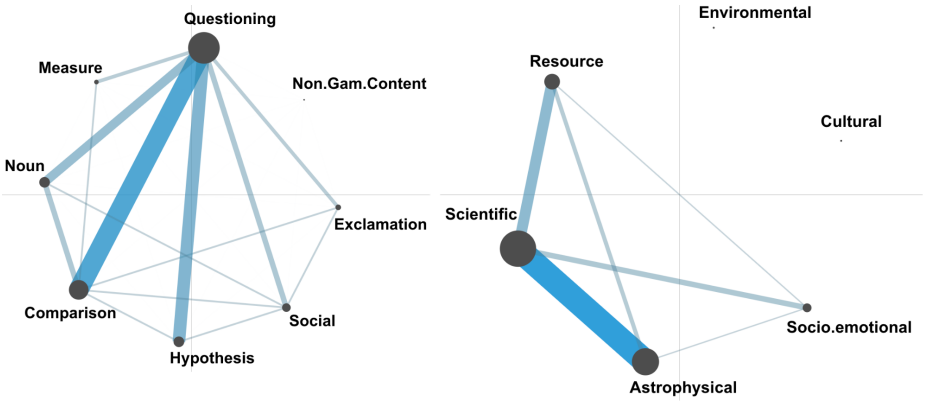



**Fig. 3.** Cat's Epistemic Network Based on Human-created Codebook (*left*) and Hybrid GPT-Human Codebook (*right*)

Meanwhile, each codebook also offers unique insights into how observation patterns differ between low- and high-interest students. The Human codebook identifies Comparison as a much more prominent posting behavior, which implies that Cat more often contrasts different elements within the game or relates game scenarios to prior knowledge than Tiger. This attentiveness to differences and connections indicates that Cat may actively use these comparisons (e.g., *"the earth on mynoan is bigger than the one on earth"*) to make sense of game dynamics. Conversely, the Hybrid codebook highlights much stronger connections between Scientific and Astrophysical (lw=0.88) for Cat, which indicates that Cat more frequently combines logical reasoning and scientific inquiry (e.g., *"why is the tides so low"*) with specific observations of planetary features (e.g., *"the moon is to aggresive to the water to make it over flow"*) in the game. The synergy reflects that Cat approaches the game in a way that is both exploratory and analytical. The Hybrid codebook also reveals a stronger connection between Scientific and Resource constructs (lw=0.39) for Cat, reflecting Cat's higher interest, as these two constructs are strongly correlated with situational interest and interest in astronomy in the previous analysis.

## 4 Discussion and Conclusion

In this study, we compare two codebooks—one created by GPT with human refinement and the other by humans—both developed from the same data source and designed to achieve the same research goal (i.e., explore how students' in-game observation patterns relate to changes in their interests). Our objective is to understand how each codebook uniquely captures and interprets these patterns through both quantitative and qualitative methods. Ultimately, we hope not only to surface these insights at the codebook level, but also to leverage them, individually or collectively, to inform strategies that can effectively foster and sustain student interest in the game.

Inspection of the two codebooks shows that they cover overlapping themes. Both codebooks each identify two kinds of observations that might be seen as off task: the Human codebook identifies *Social* and *Non-game* observations, and the Hybrid codebook identifies *Socioemotional Responses* and *Cultural & Playful References.* However, the codebooks also diverge in important ways. Specifically, the Human codebook focuses on more concrete, micro-level characteristics that reflect the kinds of scientific processes that students are asked to engage in. For example, it distinguishes between students who are excited about game elements (*Exclamations),* those who are merely naming things (*Nouns*), and those observing specific physical attributes that the game is designed to teach (*Measure/Descriptive*). It also distinguishes between those who are making these more basic observations and those who are attempting high-order scientific speech acts, such as asking *Questions*, making *Comparisons*, or generating *Hypotheses*. In some ways, the Hybrid codebook reflects a similar approach: identifying both off-task themes and then differentiating between more basic, on-task observations (*Resource Identification & Utilization* and *Astrophysical & Planetary Observation*) and those that require more scientific engagement and inference on the part of the student (*Scientific Inquiry & Reasoning* and *Environmental Adaptation & Survival Strategies*). However, in contrast to the more granular categories identified in the Human codebook, categories within the Hybrid codebook reflect a broader, higher-level approach.

The strategic differences between these two inductive codebooks are interesting, both because they correlate to different external measures and because GPT was not given a prompt that specified that these codes would be used for understanding interest development. Instead, GPT's strategic approach may stem from its training on extensive textual data, which allows GPT to recognize patterns across many contexts. Although GPT appears to deprioritize some elements of specificity and concreteness, its more abstract categories appear relevant to the current coding context.

Our results show that the Hybrid codebook's constructs have stronger correlations with measures of situational and astronomy interest. This suggests that these broader, higher-level categories may therefore serve as better indicators for designing interventions. However, the presence of different codes in the Human codebook – even if non-significant in this case -- suggests that there may be additional value if both Humans and LLMs initially brainstorm codes, and we then bring the two sets of codes together in a single codebook.

The ENA also show differences between the two codebooks. Although both codebooks make distinctions between observations that do and do not reflect an orientation towards the scientific tasks of the game, these distinctions emerge quite differently. In the human codebook, we see that the low-interest student (Tiger) may engage in more basic elements of the scientific process and even excitement about the game, but he shows fewer connections with higher order scientific practices than the high-interest student (Cat). Although the students show roughly the same number of connections in the model derived from Human codes, the patterns are distinct. In contrast, there is a clear distinction between Tiger and Cat in terms of the number of connections show in the ENA models from Hybrid codes. Here, the low-interest student (Tiger) has significantly more connections with the two off-task codes (*Socioemotional Responses* and *Cultural & Playful References*). Although the high-interest student (Cat) has some

connections with *Socioemotional Responses*, most of his observations reflect a deeper engagement with the scientific elements of the game.

The degree to which the ENA patterns may be easier to operationalize when trying to develop real-time supports remains to be seen. It is possible that the distinction between the types of categories identified in the Human codebook may provide more concrete strategies for camp facilitators who are trying to improve student interest and learning. That is, advising students to move from naming things to making comparisons may be a better strategy than telling them to make more *Astrophysical* than *Environmental* observations. However, if the facilitators only need an alert that the student exhibits indications of lower interest (rather than details about their observations), then the Hybrid codebook may suffice. Like any exploratory study, some limitations could potentially affect the conclusions that can be drawn. Since our ENA analysis examines individual cases (two students), it is important to recognize that the interpretations of other students may differ; any case-based analysis only fully applies to the cases studied, even if they were chosen to be representative of a broader set of cases. In this case, they were chosen to be extreme (students with lowest and highest interest) in order to illustrate the phenomena surrounding interest, but future research should examine whether similar patterns emerge within less extreme cases. Moreso, we are only comparing codebooks for one context and type of phenomenon – different findings may be seen in other data.

By comparing a codebook that was developed using a Hybrid Human-LLM approach to a fully-human approach, this paper contributes to the ongoing conversation on leveraging LLMs to support the analysis of behavior and experiences within quantitative ethnography. Through our findings, we hope to offer insights into the advantages of each approach individually, while highlighting the potential benefits of integrating them for more comprehensive analysis. This combination has the potential to create a richer, multidimensional understanding of student experiences, which has a range of potential applications to QE research and practice in education and other domains.

**Acknowledgments.** This research was supported by National Science Foundation Grants # DRL2301173.

## References

1. Barany, A., Nasiar, N., ... Baker, R. S.: ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. Intl. Conf. on Artificial Intel, 134–149 (2024)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. of the Royal Statistical Society: Series B (Methodological) 57(1), 289–300 (1995)
3. Boeder, J. D., Postlewaite, E. L., Renninger, K. A., Hidi, S. E.: Construction and validation of the Interest Development Scale. Motivation Science 7(1), 68–82 (2021)

4. Carpenter, D., Emerson, A., Mott, B. W., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., Lester, J. C.: Detecting off-task behavior from student dialogue in game-based collaborative learning. Artificial Intelligence in Education, 55–66 (2020)
5. Carpenter, D., Geden, M., Rowe, J., Azevedo, R., Lester, J.: Automated analysis of middle school students' written reflections during game-based learning. Artificial Intelligence in Education, 67–78 (2020)
6. Conati, C., Chabbal, R., Maclaren, H.: A study on using biometric sensors for monitoring user emotions in educational games. Workshop on Assessing & Adapting to User Attitudes & Affect (2003)
7. Dai, S. C., Xiong, A., Ku, L. W.: LLM-in-the-loop: Leveraging large language model for thematic analysis. arXiv preprint arXiv:2310.15100 (2023)
8. Eagan, B. R., Rogers, B., ... Shaffer, D. W.: Can we rely on IRR? Testing the assumptions of inter-rater reliability. Intl. Conf. on Computer Supported Collaborative Learning (2017)
9. Gadbury, M.: An exploratory analysis of the relationship between interest, knowledge, and engagement of adolescents playing Minecraft in STEM-focused summer camps. Doctoral dissertation, UIUC (2024)
10. Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J. J., Perrault, S. T.: CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. CHI Conf. on Human Factors in Computing Systems, 1–29 (2024)
11. Gupta, A., Carpenter, D., Min, W., Rowe, J., Azevedo, R., Lester, J.: Enhancing multimodal goal recognition in open-world games with natural language player reflections. AAAI Conf. on Artificial Intelligence and Interactive Digital Entertainment 18(1), 37–44 (2022)
12. Im, E. O., Chee, W.: An online forum as a qualitative research method: Practical issues. Nursing Research 55(4), 267–273 (2006)
13. Jeon, J., Lee, S.: Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. Education and Information Technologies 28(12), 15873–15892 (2023)
14. Karumbaiah, S., Baker, R. S., Shute, V.: Predicting quitting in students playing a learning game. 11th Intl. Conf. on Educational Data Mining, 167–176 (2018)
15. Lane, H. C., Gadbury, M., Ginger, J., Yi, S., Comins, N., Henhapl, J., Rivera-Rogers, A.: Triggering STEM interest with Minecraft in a hybrid summer camp. Technology, Mind, and Behavior 3(4) (2022)
16. Lee, S. J., Liu, Y. E., Popovic, Z.: Learning individual behavior in an educational game: A data-driven approach. 7th Intl. Conf. on Educational Data Mining, 114–121 (2014)
17. Li, Z., Dohan, D., Abramson, C.: Qualitative coding in the computational era: A hybrid approach to improve reliability and reduce effort for coding ethnographic interviews. Socius 7, 23780231211062345 (2021)
18. Linnenbrink-Garcia, L., Durik, A., Conley, A., Barron, K., Tauer, J., Karabenick, S., Harackiewicz, J.: Measuring situational interest in academic domains. Ed. & Psychological Measurement 70(4), 647–671 (2010)
19. Liu, X., Zambrano, A., Barany, A., Ocumpaugh, J., ... Baker, R. S.: Investigating learner interest and observation patterns in a Minecraft virtual astronomy environment. Intl. Conf. on Quantitative Ethnography, 19–34 (2024)
20. Liu, X., Zhang, J., Barany, A., Pankiewicz, M., Baker, R. S.: Assessing the potential and limits of large language models in qualitative coding. Intl. Conf. on Quantitative Ethnography, 89–103 (2024)
21. Lu, A. S., Green, M. C., Thompson, D.: Using narrative game design to increase children's physical activity: Exploratory thematic analysis. JMIR Serious Games 7(4), e16031 (2019)

22. Marathe, M., Toyama, K.: Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. Conf. on Human Factors in Computing Systems, 1–12 (2018)
23. Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., De Paoli, S.: Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? Computer Methods and Programs in Biomedicine 255, 108356 (2024)
24. Park, K., Sohn, H., Mott, B., Min, W., Saleh, A., Glazewski, K., ... Lester, J.: Detecting disruptive talk in student chat-based discussion within collaborative game-based learning environments. 11th Intl. Learning Analytics and Knowledge Conf., 405–415 (2021)
25. Perrin, A. J.: The CodeRead system: Using natural language processing to automate coding of qualitative data. Social Science Computer Review 19(2), 213–220 (2001)
26. Roberts, K., Dowell, A., Nie, J.: Attempting rigor and replicability in thematic analysis of qualitative research data: A case study of codebook development. BMC Med. Research Methodology 19(1), 1–8 (2019)
27. Rosnow, R. L., Rosenthal, R.: If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). (1991)
28. Shaffer, D. W., Ruis, A. R.: How we code. Intl. Conf. on Quantitative Ethnography, 62–77 (2021)
29. Shah, D., Kames, E., Clark, M., Morkos, B.: Development of a coding scheme for qualitative analysis of student motivation in senior capstone design. Intl. Design Engineering Technical Conf.s and Computers and Information in Engineering Conf., V003T04A005 (2019)
30. Shroff, R. H., Vogel, D. R., Coombes, J., Lee, F.: Student e-learning intrinsic motivation: A qualitative analysis. Communications of the Association for Information Systems 19(1), 12 (2007)
31. Tablatin, C. L. S., Casano, J. D., Rodrigo, M. M. T.: Using Minecraft to cultivate student interest in STEM. Frontiers in Education, 1127984 (2023)
32. Tadayon, M., Pottie, G. J.: Predicting student performance in an educational game using a hidden Markov model. IEEE Transactions on Education 63(4), 299–304 (2020)
33. Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., Monteith, B. G.: An examination of the use of large language models to aid analysis of textual data. Intl. J. of Qualitative Methods 23, 16094069241231168 (2024)
34. Wang, K. D., Cock, J. M., Käser, T., Bumbacher, E.: A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices. British J. of Educational Technology 54(1), 192–221 (2023)
35. Wei, Z., Nasiar, N., Zambrano, A. F., Liu, X., Ocumpaugh, J., Barany, A., Baker, R. S., Giordano, C.: Exploring students' interest-driven patterns of scientific observation in Minecraft. To appear in Intl. Conf. of the Learning Sciences – ISLS (2025)
36. Yi, S., Gadbury, M., Lane, H. C.: Coding and analyzing scientific observations from middle school students in Minecraft. 14th Intl. Conf. Learning Sciences (2020)
37. Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., Nasiar, N.: From nCoder to ChatGPT: From automated coding to refining human coding. Intl. Conf. on Quantitative Ethnography, 470–485 (2023)
38. Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., Carroll, J. M.: Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. arXiv preprint arXiv:2309.10771 (2023)
39. Zörgő, S., Swiecki, Z., Ruis, A. R.: Exploring the effects of segmentation on semi-structured interview data with epistemic network analysis. Intl. Conf. on Quantitative Ethnography, 78–90 (2021)