Collaborative LLMs in Academic Assessments

Miguel Fortuno

*School for Modeling, Simulation, and Training*

*University of Central Florida* Orlando, FL  
mi551225@ucf.edu

LTC Benjamin Glasgall, US Army  
*School for Modeling, Simulation, and Training*

*University of Central Florida*

Orlando, FL  
be886366@ucf.edu

David Organista   
*School for Modeling, Simulation, and Training*

*University of Central Florida*Orlando, FL  
david.organista@ucf.edu

*Abstract*—This study explores the domain of Large Language Models (LLMs) in academic assessments, addressing a gap in existing literature concerning their collective application in standardized test settings. While prior research has examined the capabilities of individual LLMs in educational contexts, the potential for multi-agent LLM collaboration remains largely unstudied, particularly in improving performance and accuracy in academic testing. This research specifically investigates the impact of collaboration between LLMs has on answering standardized test questions and evaluating performance across a variety of subjects, including US history, government, environmental science, human geography, and physics. Using a quantitative approach, the study employed three different LLM agents—OpenAI’s GPT-4, Google’s Gemini 1.0 Pro, and Anthropic’s Claude 3 Opus—in a controlled environment to simulate the test taking process. The study hypothesizes that multiple collaborative agents can improve test accuracy and efficiency compared to singular LLM agents. Thes results show that multi-agent collaboration contributes on an overall basis did not significantly improve question accuracy compared to individual agents. By analyzing performance metrics of single versus multi-agent groups, this study intends to contribute to previous works designed to assess AI-agents’ strength and weakness. Additionally, this study looks to build user confidence in LLM performance and determine potential use of multi-agent collaboration.

Keywords—Large Language Models, Academic Assessments, Educational Technology, Multi-agent Systems, Collaborative Artificial Intelligence, AI in Education, Standardized Testing

# INTRODUCTION

Large Language Models (LLMs) combine natural language processing (NLP) and vast amounts of training data with generative artificial intelligence (AI). These systems are most popularly known as *chatbot* where each *chatbot* can be referred to as an agent. For example, OpenAI’s ChatGPT and Google’s Gemini are chatbots but are referred to as agents in the context of this work [1]. As the collaboration of human and AI is explored, one focus is the measurements of accuracy of one to multiple LLM agents (i.e. ChatGPT, Bard, Meta’s LLaMA) in different contexts and domains to assess the comparability of an LLM to other LLMs and to human professionals [2], [3]. Existing research has largely focused on evaluating single LLM applications against open-ended questions as well as higher level assessment exams in the medical and legal professions [4], [5], leaving the potential benefits of LLM collaboration largely unexplored, specifically in objective standardized testing environments. This gap exists despite a growing recognition of LLMs’ ability to process and synthesize vast amounts of information, suggesting that a collaborative approach could enhance test performance [6]. To this end, this paper aims to explore two interrelated questions:

1. To what extent does the number of collaborative LLMs affect performance on college level preparatory exams across a variety of academic subjects?

2. What is the ideal number of LLM agents to work together to achieve optimal results while managing computational effort?

By exploring how multiple LLMs can work together on standardized tests, AI-focused educational research can work to establish a level of confidence in AI tools for academic settings, which has been demonstrated in the case of teacher-student LLM interactions [8]. Furthermore, by evaluating cutting-edge LLMs, this study promotes the continuation of the development of better AI tutors and thus, better students. Understanding the potential collaborative abilities of LLMs in academic testing is important now due to the increasing integration of AI into education. Our hypothesis is that the collaborative effort of multiple LLMs will improve performance on academic testing compared to a single agent, as multiple agents can leverage different knowledge bases and reasoning strategies, leading to higher accuracy, as supported by previous research into LLM collaboration [7]. In addition, accuracy will improve as LLMs leverage each other's test-taking strengths and weaknesses. The results were compared by accuracy level across the different protocols to evaluate the efficacy of collaboration of LLMs.

# BACKGROUND

The context of our research problem targets the newly evolving field by which, instead of utilizing a human to LLM model, LLM-to-LLM interactions are the main goal for target topics such as problem solving. Researchers have delved into areas such as enriching lower scale LLMs with a larger model’s advantage to trainable outputs in a “teacher-student” collaboration [8], [9], and utilizing LLMs of the same family and different families to simulate debates on selected topics to attempt a consensus [7]. The avenue of LLM-to-LLM interactions can open new prospectives in the field of AI and can potentially further enhance the accuracy of LLMs through covering the weaknesses of individual LLMs.

A review of the literature showed common themes involving validating the accuracies of LLMs against one another or against other model/human evaluation frameworks [1], [2], [3], [5],[6], [10-16], and utilizing LLM’s to enrich other model frameworks and datasets, improving their accuracies to higher levels than without LLM support [17], [18]. Additionally, regarding LLM-to-LLM collaboration, the concept of utilizing LLMs to work together in test taking does not seem present in the current literature. We found that current works in the LLM-to-LLM collaboration focused specifically on performance enhancement and collaborative works outside of test taking. For example, [8] looked at performance accuracies in a “teacher-student” LLM interaction, with larger LLM’s primarily playing the teacher role and smaller LLM’s the student role. and found that teacher LLMs can improve the performance of student LLMs in chain-of-thought reasoning, as well as decreased the performance if the teacher LLM is purposefully feeding misinformation. [9], subsequently, highlighted a more argumentative approach on LLM-to-LLM interaction via several formats of debates, focusing on a simplistic form of argumentative reasoning. The authors gave several examples of the type of prompts they utilized for the different LLMs, as well as an established debate framework. This debate framework allowed for LLMs to collaborate and come to a consensus with each other, which served as a basis for our research problem for issues concerning consensus on multiple choice questions.

The exploration of LLM-to-LLM interactions within a synergistic collaboration framework [19] presents a novel avenue for leveraging LLMs’ computational potential. AI agents, as described in [20], embody human-like autonomous entities with specific objectives, offering a foundation for our methodology. By maximizing the collaborative framework, we aimed to refine the process of subject mastery, enhancing LLMs ability to communicate with one another and self-improve its reasoning on standardized testing [21]. These agents, by simulating human judgment, facilitated a dynamic collaborative environment where LLMs not only self-improve their subject mastery and decision-making but also provide insights to other LLMs that guide further exploration and synergy [18],[22]. This approach underscored the potential of AI models as educators that guide decision-making through informed recommendations for the future of education [23].

Existing research has primarily focused on the direct interaction within the same LLM without considering the integration of multi-agent collaboration. However, these studies demonstrate the feasibility of LLM-to-LLM collaboration for task resolution and the generation of consensus through debate [22], yet they overlook the critical aspect of subject mastery and the dynamic inclusion of role differentiation for a synergetic non-linear system. Our proposal seeks to bridge this gap by introducing multiple AI agents within the same environment with distinct roles, yet with the same task completion — standardized test taking. Our methodology extends beyond the scope of current literature by offering a comprehensive solution that incorporates the strengths of LLM-to-LLM interaction and the flexibility of role differentiation for real-time refinement and domain knowledge judgment and mastery. This approach promises a significant advancement in the application of LLMs for educational purposes, providing a pathway not only for users to engage with AI in a more meaningful and personalized manner, but for LLMs to improve to next echelon of collaborative behaviors.

[24] discussed how LLMs have been used to support academic writing and code generation while investigating the use of LLMs to support advanced modeling and simulation, specifically explaining model structures, summarizing outputs, enhancing accessibility, and providing explanation of errors to non-experts. The study highlighted that researchers are continuing to use LLM’s natural language processing as a tool to complete tasks previously supported by other technologies (such as internet-based search engines to retrieve information). In this study, the author focused solely on a single LLM (GPT-4) for M&S tasks support, warranting potential further investigation whether multi-agent collaboration could improve model validation and simulation support.

Additionally, [25] investigated how Theory of Mind (ToM), which is the understanding that others possess distinct knowledge and perspectives, plays a role in understanding the efficacy of LLM-to-LLM interactions. By incorporating the concept of ToM into LLMs collaboration, researchers enable them to simulate self-awareness and perspective-taking to make more informed interventions, significantly improving the learning outcomes of their counterparts [8], [9]. Furthermore, [14] demonstrated that LLM-to-LLM collaboration implies ToM capabilities through teaming to complete a scenario with a set of predefined rules set in place by the authors. These studies demonstrated different applications of LLM-to-LLM collaboration. However, while they helped broaden the emerging field of LLM collaboration, a review of the literature shows no papers discussing utilizing multiple LLMs to gauge improvements in accuracy of test taking compared to a single LLM. Given this gap in the literature, our study sought to determine the impact of multi-agent collaboration on test taking performance and accuracy as well as to provide a path for researchers to develop future research.

# METHODOLOGY

This research project aimed to investigate the impact of collaborative LLM agents on accuracy and efficiency of responses in academic assessments, specifically within the context of college preparation examinations across a variety of subjects, i.e. US history, US government and politics, environmental science, human geography, and physics. These subjects were chosen based on an examination of the five lowest averages from high school students taking college-level preparation exams [26]. Grounded in the premise that diversity in problem-solving approaches enhanced answer outcome quality, this study hypothesized that a collective effort between multiple LLM agents, i.e., OpenAI’s GPT-4 (GPT), Anthropic’ s Claude 3 Opus (Claude), and Google’s Gemini 1.0 Pro (Gemini), will improve performance compared to individual LLM operations. These LLMs were chosen based on a variety of criteria. GPT-4 was selected as it has been proven to perform better than other LLMs, specifically its predecessor GPT-3.5 and Bard, while evaluating reasoning abilities through question-and-answer tasks [27], [28]. Gemini and Claude were chosen as the alternative comparative models due to their recent emergence as competitive agents to GPT-4 in the closed source technology space. This hypothesis is supported by previous research indicating that collaboration between different agents leads to improved problem solving [7], [18].

Unlike previous studies that have primarily focused on assessing the capabilities and limitations of single LLM agents for higher education and vocational exams [4],[5], this project proposed an experimental approach that evaluates the combined strengths of different agents. By creating diverse “teams” of LLMs and comparing their performance to that of individual agents on the standardized portion of college-level preparatory exam, this study aimed to gain insight into the potential of collaborative AI to improve educational assessments and learning outcomes. This approach not only expanded on the current understanding of Artificial Intelligence (AI)’s role in education [8], [9], but also presents opportunities to explore applications of agent-assisted instruction such as AI intelligent tutoring.

## Approach

This study used a quantitative research approach to evaluate the efficacy of collaborative to accurately answer test questions, leveraging the APIs (Application Programming Interface) of each of the LLMs using an integrated development environoment. Each exam was structured into a digital format compatible with the input requirements of each LLM’s API. The experiment had a dual focus on both the total aggregate accuracy as well as specific academic and analytical skills tested by the exam. Based on the examination assessment keys, the academic skills assessed by the tests are causation, contextualization, continuity and change over time, analyzing evidence, and comparison (see Table 1 for the distribution of academic skills assessed). This dual focus allowed for a nuanced understanding of the agents’ effectiveness not just in terms of raw scores but also their ability to handle various types of categorically-defined questions.

By systematically analyzing the performance of different LLM configurations—individual agents versus collaborative groups comprising of GPT, Gemini and Claude - this study sought to quantify the impact of model collaboration on educational assessment. The integration and utilization of the various LLMs’ APIs enabled real-time interaction with each model to simulate the exam-taking process under controlled conditions. A conceptual model of our LLM collaboration accounting for one vs. two vs. three “students” is showcased in Figure1.

A diagram of a diagram

Description automatically generated

Figure 1: High level model depicting unique agent grouping.

## Conditions/Scenario

This study employed an integrated development environment (i.e., Jupiter Notebook) using Google Colaboratory (Colab) to assess the performance of three listed LLM agents in answering each set of preparatory exam questions. The simulation was structured to replicate the conditions of an actual exam, with the objective of evaluating the LLMs’ abilities across five key categorically defined skills: contextualization, analyzing evidence, continuity and change over time, and comparison. The simulation presented each LLM combination with the same set of questions, administered through LLMs’ respective APIs, as a baseline to determine the accuracy of single-LLM testing and evaluate the strengths and weaknesses of each respective agent. Once that is determined, the next step was the collaboration phase as defined in Figure 2.

A diagram of a flowchart

Description automatically generated

Figure 2: 3-Round Multi-agent collaboration process with arbitration.

During the collaboration phase, the LLMs were grouped together in either 2 or 3-agent combinations and made to go through a 3-round system to ensure answer agreement. In round 1, the individual LLMs in a group were given each individual question from each subject and asked to provide a confidence level (0-100%, where 0 is least likely and 100 is most likely to be the correct answer) to their answer as well as a one-sentence explanation for why it chose its answers. The letter choice outputs were then compared to determine agreement. If the LLMs did not agree on an answer, they underwent a second round. In this round, they were given the question again, along with each LLM’s first-round outputs—comprising letter choice, confidence level, and explanation—as input, thus creating a feedback loop. If the group disagreed after the second round, a final round was made with an arbiter (judge) agent. For two-LLM groups, the judge agent was the LLM that is not part of the 2-LLM group and thus not present in the first 2 rounds. For the case of all three LLMs, the judge was determined by the greatest overall strength among all five subjects. In both cases, the judge LLM took each agent’s round 2 outputs and a report on each agent’s skill strengths and weaknesses including its own to decide on the final outcome for that question. The simulation ultimately measured how these collaborative interactions influenced the correctness of the LLMs’ group responses compared to their individual outputs using an answer key with correct letter choices and skilled assessed. By mimicking a real-world collaborative learning environment, the study evaluated the potential of multi-agent models versus individual agent models and to what extent their performance could be affected.

*Table 1:Skill Distribution of Questions*

|  |  |
| --- | --- |
| **Skill Assessed** | **Number of Questions** |
| Contextualization | 106 |
| Causation | 121 |
| Analyzing Evidence | 140 |
| Continuity and Change Over Time | 56 |
| Comparison | 65 |
| Total | 488 |

## Procedures Summary

The experimentation involved first setting up each LLM via their API. All code was generated using the Python programming language. The answer keys were generated and used only for comparing the LLM’s choice answer to the true and correct answer for each question. Several questions contained images as part of the question or answer choices, with environmental science having the least number of image-based questions and physics having the most, making up approximately 25% of the question dataset. These images were converted into text descriptions using GPT-4 Vision and replaced by the images during the test preparation.

For individual agent testing, agents were prompted to provide just the question number and alphabetical answer choice. The prompt was then updated for the agents to include a confidence level and an explanation during multi-agent collaboration. Each run included the completion of a full test subject by one combination group. Following each exam run, the agents’ test answers were compared against the answer key to obtain the test accuracy for that run of code. The accuracy count was output into a comma-separated values (CSV) file for each unique LLM group, with multi-agent groups outputting how many rounds it took to arrive at the agreed upon answer. Each round’s answers were auto graded to see if the agreement resulted in a correct answer. Disputed answers were marked as incorrect, even if at least one agent was correct.

## Metrics

The primary metrics collected during this research were overall test accuracy and question skill accuracy of individual and collaborative LLM groups. Overall test accuracy is defined as the ratio of total correct responses over the total exam questions. Question skill accuracy is defined as the ratio of total correct responses over the total exam questions for a skill field. Overall accuracy has been the statistical metric used to evaluate LLMs against each other or human subjects, assessing the accuracy of an LLM [1], [3]. Question skill accuracy is considered to understand the potential strengths and weaknesses of each LLM. It is understood that LLMs can be better performers in particular domain fields compared to others [1], [3]. By analyzing the skill sets from the respective examinations, we were able to make insights into the categories that the selected LLMs consistently excel at and whether multi-LLM collaboration provided performance impacts for LLMs with smaller training data.

## Procedures

The detailed procedure steps of our experimentation can be defined as followed:

1. Initialize the code environment and insert question and answer key files.
2. Insert APIs for each agent.
3. Run the code program for each unique LLM group; group experimentation is as followed:
4. GPT-4, Claude 3 Opus, and Gemini as individual groups.
5. Three Two-LLM unique groups (GPT4 and Gemini, GPT4 and Claude 3 Opus, Claude 3 Opus and Gemini)
6. One Three-LLM unique group
7. Collect multiple-choice answers from each LLM group.
8. Compare the overall test accuracy and question skill accuracy for each LLM group to the answer keys generated per subject.
9. For multi-agent groups, gather the number of rounds it took to reach concurrence.
10. Calculate descriptive statistics to determine any significant differences between all agent groups (i.e. One-Way ANOVA or Kruskal-Wallis Test and Tukey’s HSD test or Dunn’s test, based on normality of answers).

# AGENT RESULTS

## Count

Each agent group received questions from each subject to obtain an average across all subjects. A total count across all subjects from each agent group were captured to assess overall question count accuracy. The total correct answer count across all agent groups is listed in Table 2.

Table 2: Correct Answer Count by Subject

| Agent Group | Subject | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| US His. | US Gov. | Phys. | Hum. Geogr. | Environ. Sci. | Total |
| GPT | 54 | 81 | 36 | 83 | 145 | 399 |
| CL | 54 | 81 | 46 | 86 | 148 | 415 |
| GEM | 49 | 70 | 24 | 73 | 133 | 349 |
| GPT – CL | 52 | 86 | 42 | 85 | 147 | 412 |
| GPT – GEM | 50 | 87 | 36 | 81 | 146 | 400 |
| CL – GEM | 51 | 84 | 35 | 78 | 145 | 393 |
| ALL | 54 | 86 | 43 | 80 | 146 | 409 |

From Table 2, the results showed that Gemini had the overall least number of correct answers among single agent groups; Claude had the most overall correct answers among the single agent groups. When it comes to the multi-agent groups, the GPT and Claude group had the second highest number of correct answers among all groups, with all three agents highlighting the third highest number of correct answers. Figure 3 presents a visual representation of the total correct answers by agent group across all subjects.

A graph of numbers and colors

Description automatically generated

Figure 3: Bar graph displaying the total number of correct responses for each agent group.

After determining the correct answer count for each subject, the average and standard deviation were obtained for each agent group. The results for each statistic are displayed in Table 3 and 4, respectively.

Table 3: Agent Group Averages

| Agent Group | Subject | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| US His. | US Gov. | Phys. | Hum. Geogr. | Environ. Sci. | Overall Avg. |
| GPT | 98.2% | 84.4% | 48.0% | 79.0% | 92.4% | 80.4% |
| CL | 98.2% | 84.4% | 61.3% | 81.9% | 94.3% | 84.0% |
| GEM | 89.1% | 72.9% | 32.0% | 69.5% | 84.7% | 69.6% |
| GPT – CL | 94.5% | 89.6% | 56.0% | 81.0% | 93.6% | 82.9% |
| GPT – GEM | 90.9% | 90.6% | 48.0% | 77.1% | 93.0% | 79.9% |
| CL – GEM | 92.7% | 87.5% | 46.7% | 74.3% | 92.4% | 78.7% |
| ALL | 98.2% | 89.6% | 57.3% | 76.2% | 93.0% | 82.9% |

The range of averages among the groups are as followed:

* US History: 90.9% – 98.2%
* US Government: 72.9% – 90.6%
* Physics: 32.0% - 61.3%
* Human Geography: 69.5% - 94.3%
* Environmental Science: 84.7% - 94.3%

From the range of averages, physics had the largest spread among the agent groups at 29.3%, followed closely by human geography with a spread of 24.8%. US History had the least spread among the agent groups at 7.3%.

The highest and lowest averages among the subject topics, by agent group, are as followed:

* US History: Highest – GPT-4 (individual), Claude (individual), & all three agents; Lowest – Gemini
* US Government: Highest – GPT-4 & Gemini grouping; Lowest – Gemini
* Physics: Highest – Claude; Lowest – Gemini
* Human Geography: Highest - Claude; Lowest – Gemini
* Environmental Science: Highest – Claude; Lowest – Gemini

Figure 3 depicts the average clusters for each subject among the agent groups.

A graph of different colored bars

Description automatically generated

Figure 4: Graph displaying the average spread per subject.

Table 4: Agent Group Standard Deviation

| Agent Group | Subject | | | | |
| --- | --- | --- | --- | --- | --- |
| US His. | US Gov. | Phys. | Hum. Geogr. | Environ. Sci. |
| GPT | 13.5% | 36.5% | 50.3% | 40.9% | 26.7% |
| CL | 13.5% | 36.5% | 49.0% | 38.7% | 23.3% |
| GEM | 31.5% | 44.7% | 47.0% | 46.3% | 36.1% |
| GPT – CL | 22.9% | 30.7% | 50.0% | 39.5% | 24.5% |
| GPT – GEM | 29.0% | 29.3% | 50.3% | 42.2% | 25.6% |
| CL – GEM | 26.2% | 33.2% | 50.2% | 43.9% | 26.7% |
| ALL | 13.5% | 30.7% | 49.8% | 42.8% | 25.6% |

Figure 5 provides a visual representation of the individual agents across the skills. The figure shows that Claude overall had the highest accuracies for the five assessed skills while Gemini had the least overall accuracy.

A graph showing different colored and green rectangles

Description automatically generated

Figure 5: Breakdown of Test skill assessed by individual agents.

## Statistical Analysis

Table 5 showcases the Kruskal-Wallis H test statistic among the five subjects. The Kruskal-Wallis test was chosen due to the singular run of each subject question dataset for each agent group.

Table 5: Kruskal-Wallis Test Results

| Subject | Kruskal-Wallis Test Results |
| --- | --- |
| H Test Statistic |
| US His. | 9.14 |
| US Gov. | 17.42\*\* |
| Phys. | 16.81\* |
| Hum. Geogr. | 6.38 |
| Environ. Sci. | 13.29\* |

\*\*α=0.01; \*α=0.05

There were no statistical differences among the groups for US history and US geography while physics, environmental science, and US government had statistical significance for group differences. Environmental science and physics were significant at the alpha level of 0.05 and US government was significant at the alpha level of 0.01. Table 6 depicts those groups that have statistically significant differences at the alpha level of 0.05 after calculating the p-valve from Dunn’s test. Environmental science and US Government had six p-values while physics had three p-values that met the alpha level criteria.

Table 6: Dunn's Test P-Value Results

| Group Cf. | P-Values by Subject | | |
| --- | --- | --- | --- |
| Environ. Sci.. | US Gov. | Phys. |
| GPT/ GEM | 0.0132 | 0.0242 |  |
| CL/ GEM | 0.0020 | 0.0242 | 0.0003 |
| GEM/ GPT – CL | 0.0038 | 0.0010 | 0.0033 |
| GEM/ GPT – GEM | 0.0073 | 0.0005 |  |
| GEM/ CL – GEM | 0.0132 | 0.0041 |  |
| GEM/ALL | 0.0073 | 0.0010 | 0.0019 |

Table 6 shows that all the significant differences resulting from Dunn’s test were associated with Gemini. For environmental science, the comparison between Claude and Gemini generated the smallest p-valve. For US government, the comparison between Gemini and the multi-agent group consisting of GPT and Gemini generated the smallest p-value. For physics, the comparison between Claude and Gemini generated the smallest p-value.

# Discussion

There were mixed results depending on the topic of subject, from positive improvements in US Government to neutral and negative improvements. However, the results obtained were similar to the results obtained from the companies’ technical reports [28 – 30].

When looking at the subject of physics, for example, the Anthropic team reports an accuracy of 60.1% for a zero-shot result on the MATH dataset, a dataset consisting of 12,500 problems from seven different mathematical topics [31]. Google reports an accuracy of 32.6% for Gemini 1.0 Pro on a four-shot assessment through MATH [29]. While OpenAI did not utilize assessment on the MATH dataset, they report a variety of college-level topics, of which one was physics 2 [28]. While this study focused on physics 1, the results showed that GPT received a lower accuracy than what is reported, but within the performance band of improvement when comparing to GPT-3.5 [28]. This is believed to be due to the potential inaccuracies from GPT-4Vision when converting the numerous images in the physics dataset to text, possibly showing hallucinatory responses as addressed in GPT-4’s technical report [28]

When addressing multi-agent groups, there is one trend that was apparent among the five subjects. All four multi-agent groups answering US government test questions performed visually better than their individual counterparts. However, statistical differences were primary tied to Gemini’s results to other group results. GPT-4 and Claude did not have statistical differences between each other or the multi-agent groups.

When removing Gemini from the observations, none of the two-agent groups performed better than the individual assessments in Human Geography and US History. These lower performances were not statistically significance when comparing the differences in the averages among groups.

When addressing the research question, the following are notated:

*To what extent does the number of collaborative LLMs affect performance on college level preparatory exams across a variety of academic subjects?*

As addressed above, US government question averages improved from the individual counterparts, but these results are statistically significant only for Gemini and not for GPT-4 and Claude 3. Gemini benefitted the from multi-agent collaboration compared to the other individual agents, having a 20-25% improvement in the average while working together with GPT-4 and Claude 3. This is highlighted by the lowest p-value when conducting Dunn’s Test being the comparison between Gemini and the multi-agent group consisting of GPT-4 and Gemini.

It is important to note that LLMs generate results based on their training data. Although different LLMs are trained on different content and size of data, the consistency of higher performances on US government by collaborative groups is significant. Consistency can be explained by 2 main reasons.

1. Our current LLMs are trained by text corpora including historical and official documents and laws that are presented as highly structured “LLM friendly” information [32].

2. The datasets on which they are trained are “US-flavored” where not only is English the primary training language, but the American demographics are also overrepresented [33].

The other subjects display an overall different picture for multi-agent collaboration. On US history, GPT-4 and Claude 3 performed better than the multi-agent groups but there is no statistical significance among any of the groups. Similarly with physics, Claude 3 performed better individually, and GPT-4 has no change to its performance when paired with Gemini; these results did not depict statistical significance. Gemini shows a statistical difference when compared to Claude 3 and the multi-agent group consisting of GPT-4 and Claude 3, but not with the groups Gemini collaborated with to answer physics questions.

Human Geography repeats the pattern of physics, with Claude 3 performing the best individually, but not by a significant margin to the next best group. This leaves Environmental Science, for which all statistical differences point to the discrepancy in accuracies between Gemini and the other agent groups. While not as significant as the ones in US government, Gemini still meets an alpha level of 0.05 when placed in a collaborative group with Gemini and Claude 3. With US government, this suggests Gemini benefitted from working together with the agents used for this study.

It is also noteworthy to consider that as the rounds increase in multi-agent collaboration, so does LLM agreement and accuracy. This means that iterative processes of collaboration yield favorable results in the context of other LLM tasks where the social aspect of collaboration and decision-making is essential. This synergy may not have been leveraged for standardized testing and the potential is significant.

When looking at whether collaborative agents make a difference, our results indicate that overall, with the subjects that were chosen to conduct this experimentation, there is little to no significance, particular for GPT-4 and Claude 3. For Gemini, the overall lower performance is explained by the fact that Gemini 1.0 Pro was the only model at the time of this research that was available to the public from the Gemini family of models. Based on Google’s technical report [29], Gemini 1.0 Ultra could have proven to be comparative to Claude 3 and GPT-4 in this study and may have boosted the performance accuracies of the groups that Gemini was a part of.

This leads to the second part of the research question:

*What is the ideal number of LLM agents to work together to achieve optimal results while managing computational effort?*

Based on the average observations and statistical analysis, there is overall little to no difference for GPT-4 and Claude 3, and the multi-agent groups with which they were working. The individual performances of Claude 3 gave the highest averages and overall correct answer count. This agent is best used individually across the subjects that were analyzed.

For the case of GPT-4, the agent performed at an equivalent or higher average on US history and human geography compared to multi-agent group that contained GPT in them; for the other subjects, multi-agent groups added up to 8% to GPT’s performance , but more often the performance addition was about 1-3%. Considering the miniscule performance affects in multi-group settings, GPT-4 should be ideally used by itself to conduct performance assessments. Furthermore, GPT-4’s collaborative weakness can be highlighted by its report [28] where it may be hindered by not learning from its experience due to its pre-trained nature.

Gemini 1.0 Pro benefitted the most from multi-agent collaboration in physics, environmental science, and US government. The agent added an additional 25.3% performance boost when working together with GPT-4 and Claude-3 on the physics questions, 17.7% when working with GPT on US government questions, and 8.3% when working with GPT in a two-agent group or with both GPT and Claude 3 in a three-agent group. Therefore, Gemini ideally could achieve optimal results in these subjects when performing in the two and three agent groups but at the expense of other agents.

When it comes to the overall assessment of the ideal number of agents, based on the explanations, one agent is the ideal number to perform question-answering for the five topics. Furthermore, individual LLMs should be assessed when completing simple tasks to best leverage their strengths.

Lastly, our hypothesis for the positive improvement of multi-agent collaboration compared to individual agents can be addressed. The results for US government are in favor of the hypothesis, but only for Gemini based on the statistical analysis. For all other subjects, the hypothesis cannot be accepted based on the results obtained. For an overall assessment, the hypothesis is rejected that multi-agent collaboration leads to better averages than individual agents.

# Conclusion

This study examined the importance of considering whether multi-agent systems can positively affect the accuracies of subjects from their individual counterparts. The overall consensus is that individual agents are the most optimal for conducting question-answering assessments and similar one-shot, simple tasks. The high accuracies obtained by the individual agents for 4 of the 5 subjects highlight significant improvements in the AI field. However, care is to be taken in ensuring that these agents are used as an enhancement to learn rather than a replacement. It is crucial that educators define clear boundaries and purposes for the integration of LLMs in educational settings. LLMs should support curricular goals by enriching the content and interactivity of lessons, providing additional resources for students to explore, and offering tailored feedback to help refine their skills and understanding.

There are several ways to enhance the methodologies of this study, which could lead to substantial improvements in research outcomes. This study looked at only proprietary agent models; a comparison to open-source models, such as Meta’s LLaMa-2 and the ever-growing number of agents from the Hugging Face community, could highlight performance differences between open-source and proprietary models. Additionally, this study focused on subjects which happened to be social sciences and mathematics focused; seeing LLM collaboration for question-answering on subjects such as humanities and languages could expand the use cases of multi-agent systems. Multiple trials of the questions would help to reduce potential bias in agent assessments due to chance. Lastly, the round-base system can be improved through a more robust voting schema, such as rank choice or response-weighting.

##### References

[1] M.Z. Naser et. al., “Evaluating the Performance of Artificial Intelligence Chatbots and Large Language Models in the FE and PE Structural Exams,” *Pract. Periodical Structural Des. Construction*, Volume 29, Issue 2, Accessed: January 29, 2024, doi: 10.1061/PPSCFX.SCENG-1369. [Online]. Available: https://ascelibrary.org/doi/full/10.1061/PPSCFX.SCENG-1369

[2] Z. Elyoseph, I. Levkovich, and S. Shinan-Altman, “Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public”, *Family Med. Community Health*, Volume 12, Suppl 1, Jan. 2024, Accessed January 30, 2024, doi: 10.1136/fmch-2023-002583 [Online]. Available: <https://fmch.bmj.com/content/12/Suppl_1/e002583>

[3] A. Suάrez et al., “Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery,” *Comput. Struct. Biotechnol.*, vol. 24, pp. 46-52, Dec. 05, 2023, doi: 10.1016/j.csbj.2023.11.058. [Online]. Available: https://www.csbj.org/article/S2001-0370(23)00473-7/fulltext

[4] G. Pinto, I. Cardoso-Pereira, D. Monteiro, D. Lucena, K. Gama, “Large Language Models for Education: Grading Open Ended Questions Using ChatGPT,” in Proc. Of the XXXVII Brazilian Symposium on Software Engineering (SBES ’23), Sep. 2023, pp.293-302, doi: 10.11445/3613372.3614197

[5] M. C. Schubert, W. Wick, V. Venkataramani, "Performance of Large Language Models on a Neurology Board–Style Examination," JAMA Network Open, vol. 6, no. 12, e2346721, Dec. 2023, doi:10.1001/jamanetworkopen.2023.46721.

[6] F. Kamalov, D. Santandreu Calonge, and I. Gurrib, "New era of artificial intelligence in education: Towards a sustainable multifaceted revolution," *Sustainability*, vol. 15, no. 16, p. 12451, 2023.

[7] K. Xiong, X. Ding, Y. Cao, T. Liu, and B. Qin, "Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, Dec. 2023, pp. 7572–7590, doi: 10.18653/v1/2023.findings-emnlp.508.

[8] S. Saha, P. Hase, and M. Bansal, “Can Language Models Teach Weaker Agents? Teacher Explanations Improve Students via Personalization,” presented at the 37th Annu. Conf. Neural Inf. Process. Syst., New Orleans, LA, Dec. 10-16, 2023, Accessed: February 10, 2024, [Online]. Available: <https://neurips.cc/virtual/2023/poster/72111>

[9] Z. Abbasiantaeb, Y. Yuan, E. Kanoulas, and M. Aliannejadi, "Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions," arXiv preprint arXiv:2312.02913.

[10] Z. He *et al.*, "Large Language Models as Zero-Shot Conversational Recommenders," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage*., 2023, pp. 720-730, doi: 10.1145/3583780.3614949.

[11] H. Fan, X. Liu, J. Y. H. Fuh, W. F. Lu, and B. Li, "Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics," *J. Intell. Manuf.*, 2024. [Online]. Available: <https://doi.org/10.1007/s10845-023-02294-y>

[12] L. Zheng *et. al.*, “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,” presented at the 37th Annu. Conf. Neural Inf. Process. Syst., New Orleans, LA, Dec. 10-16, 2023, [Online]. Available: <https://openreview.net/forum?id=uccHPGDlao>

[13] J. de Curtò et al., “LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments.”, *Electron. (Basel).*, vol. 12, issue 13, Jun. 2023, Accessed February 02, 2024, doi: 10.3390/electronics12132814 [Online]. Available: <https://www.mdpi.com/2079-9292/12/13/2814>

[14] H. Li *et al.*, “Theory of Mind for Multi-Agent Collaboration via Large Language Models,” in *Proc. 2023 Conf. Empirical Methods Natural Lang. Process.*, 2023, pages 180–192. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.13.pdf>

[15] S. Arulmohan, M-J. Meurs, and S. Mosser, “Extracting Domain Models from Textual Requirements in the Era of Large Language Models”, in *2023 ACM/IEEE Int. Conf. Model Driven Eng. Syst. Companion*, Västerås, Sweden, 01-06 Oct., doi: 10.1109/MODELS-C59198.2023.00096.

[16] X. He, H. Gao, J. He, and C. Sun, “Evaluation of Large Scale Language Models on Solving Math Word Problems with Difficulty Grading,” in *2023 Int. Conf. Intell. Educ. Intell. Res.*, Wuhan, China, 05-07 Nov., doi: 10.1109/IEIR59294.2023.10391224.

[17] I. de Zarzà *et al.*, “LLM Multimodal Traffic Accident Forecasting”, *Sensors*, vol. 23, issue 22, pp. 9225, Nov. 2023, doi: 10.3390/s23229225. [Online]. Available: [https://www.mdpi.com/1424- 8220/23/22/9225](https://www.mdpi.com/1424-%208220/23/22/9225)

[18] N. Nascimento, P. Alencar and D. Cowan, "Self-Adaptive Large Language Model (LLM)-Based Multiagent Systems," in *2023* *IEEE Int. Conf. Autonomic Comput. Self-Organizing Syst. Companion*, Toronto, ON, Canada, pp. 104-109, doi: 10.1109/ACSOS-C58168.2023.00048.

[19] P. Panzarasa, N. R. Jennings, and T. J. Norman, "Formalizing collaborative decision-making and practical reasoning in multi-agent systems," *J. Logic Comput.*, vol. 12, no. 1, pp. 55-117, 2002.

[20] J. Insa-Cabrera, D. L. Dowe, S. Espana-Cubillo, M. V. Hernández-Lloreda, and J. Hernández-Orallo, "Comparing humans and AI agents," in Proc. 4th Int. Conf. Artif. General Intell. (AGI), Mountain View, CA, USA, 2011, pp. 122-132.

[21] Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2022). Training Language Models with Language Feedback. arXiv preprint arXiv:2204.14146.

[22] G. Dong et al., “Revisit Input Perturbation Problems for LLMs: A Unified Robustness Evaluation Framework for Noisy Slot Filling Task,” in *CCF Int. Conf. Natural Lang. Process. Chinese Comput.*, in Lecture Notes in Artificial Intelligence, in Lecture Notes in Computer Science, vol. 14302, Oct. 2023, pp. 682-694. [Online]. Available: <https://link.springer.com/chapter/10.1007/978-3-031-44693-1_53>

[23] T. Wu *et al*., "A brief overview of ChatGPT: The history, status quo and potential future development," in *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1122-1136, May 2023.

[24] P. J. Giabbanelli, "GPT-Based Models Meet Simulation: How to Efficiently use Large-Scale Pre-Trained Language Models Across Simulation Tasks," 2023 Winter Simulation Conference (WSC), San Antonio, TX, USA, 2023, pp. 2920-2931, doi: 10.1109/WSC60868.2023.10408017.

[25] A. Liu, H. Zhu, E. Liu, Y. Bisk, and G. Neubig, "Computational Language Acquisition with Theory of Mind," presented at the 11th Int. Conf. Learn. Representations (ICLR), Kigali, Rwanda, May 01-05,2023. [Online]. Available: <https://openreview.net/forum?id=C2ulri4duIs>

[26] College Board®. 2023. “Student Score Distributions.” Accessed 11 March 2024. <https://apcentral.collegeboard.org/media/pdf/ap-score-distributions-by-subject-2023.pdf>

[27] J. L. Espejel, E. H. Ettifouri, M. S. Y. Alassan, E. M. Chouham, and W. Dahhane, "GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts," *Natural Language Processing Journal*, vol. 5, p. 100032, 2023.

[28] OpenAI, “GPT-4 Technical Report,” 2023, arxiv:2023.08774v6

[29] Gemini Team, Google, “Gemini: A Family of Highly Capable Multimodal Models,” 2024, arxiv:2312.11805v2

[30] Anthropic, “The Claude 3 Model Family: Opus, Sonnet, Haiku,” 2024, [Online]. Available: <https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf>

[31] D. Hendrycks *et al.*, “Measuring Mathematical Problem Solving With the MATH Dataset,” presented at the 35th Annu. Conf. Neural Inf. Process. Syst. Track Dataset Benchmarks, 2021, [Online]. Available: <https://openreview.net/forum?id=7Bywt2mQsCe>

[32] M. Domingo and F. Casacuberta, "Modernizing historical documents: A user study," *Pattern Recognition Letters*, vol. 133, pp. 151-157, 2020.

[33] Buttrick, N. (2024). Studying large language models as compression algorithms for human culture. *Trends in Cognitive Sciences*.