Introduction-

Through the use of artificial intelligence (AI) and machine learning (ML), current large language models (LLMs) have demonstrated a strong ability to accurately process and generate natural language to a wide array of inputs. One such model is OpenAI’s Chat Generative Pre-Trained Transformer (ChatGPT) powered by GPT-3.5. Past studies have evaluated GPT-3.5’s performance to assess baseline competency on both the United States Medical Licensing Exam (USMLE) and other residency board exams. In the field of orthopaedic surgery, prior studies have analyzed performance on the Orthopaedics In-Training Exam (OITE) and found that GPT-3.5 performed at the level most comparable to an orthopaedic surgery intern. While previous studies have demonstrated a baseline competency of the model, further research is necessary to assess the performance of GPT-3.5 at a granular level. The current study examines GPT-3.5’s performance on the American Academy of Orthopaedic Surgeons (AAOS) self-assessment exam questions by orthopaedic subcategory.

Methods-

The AAOS offers a question bank of 3,450 questions to help prepare orthopaedic surgery residents for their board exams. The question bank is separated by subcategories shown in Table 2. Without replacement, one hundred questions were randomly sampled from all categories with at least 100 questions. All questions were sampled in categories with less than 100 questions (Miscellaneous and Anatomy Imaging). A total of 1,111 questions were sampled from this methodology. Questions were individually given to GPT-3.5 using a Python 3.8.1 coding script that utilized GPT-3.5’s Application programming interface (API). GPT-3.5’s answers were recorded and then manually graded against the AAOS answer key. Furthermore, questions were stratified by question style. Questions that were purely text-based were classified as Type 1. Questions with an associated image were classified as Type 2. GPT-3.5 refused to answer some image-associated questions due to its inability to process image inputs. These questions were noted, removed from GPT-3.5’s performance analysis, and designated as Type 3. The model’s performance was compared to a cohort of 4,496 orthopaedic residents around the country who completed the 2022 OITE. A one-way analysis of variance (ANOVA) across subcategories was performed on Type 1 question, Type 2 question, and the composite using SciPy in Python. Additionally, a chi-squared test was performed between Type 1 and Type 2 questions within a subcategory using Microsoft Excel. Significance was determined to be p < 0.05.

Results-

Table 1 outlines GPT-3.5’s performance by the 12 subcategories outlined on the AAOS question bank. For every subcategory, GPT-3.5 correctly answered a higher percentage of Type 1 questions compared to Type 2 questions. For Type 1 questions, GPT-3.5 performed the best on Adult Spine (66.0%) and the worst on Sports Medicine (47.2%), whereas for Type 2 questions the model performed best on the Miscellaneous section (50.0%) and worst on Anatomy Imaging (30.3%). For the composite (Type 1 and Type 2 questions), GPT-3.5 performed the highest on Basic Science and Adult Spine (55.0%) and the lowest on Anatomy Imaging (33.3%). GPT-3.5 refused to answer 26 of the 533 Type 2 questions (4.9%). No Type 3 questions were observed in the Adult Spine, Basic Science, and Hand/Wrist. Musculoskeletal Tumors and Disease had the most type 3 questions of any subcategory (10). Figure 1 charts GPT-3.5’s performance by subcategory in comparison to national orthopaedic surgery residency scores by postgraduate year (PGY) from PGY1 to PGY5. On the cumulative exam, residents scored 54.7%, 61.3%, 67.7%, 71.3%, and 73.2% from PGY1 to PGY5, respectively. There was no statistically significant difference found in performance between Type 1 and Type 2 questions for a given subcategory except for Adult Spine (p = 0.0381) and Pediatrics (p = 0.02799). The one-way ANOVA failed to demonstrate a correlation between GPT-3.5 performance and subcategory for Type 1, Type 2, and composite questions (p = 0.89471, p = 0.89834, and p = 0.461997 respectively).

Conclusion-

ChatGPT-3.5 demonstrated no relationship between performance and question subcategory. While only statistically significant in two subcategories (Adult Spine and Pediatrics), the fall in performance from Type 1 to Type 2 questions demonstrates the importance of image processing capabilities needed to correctly answer board-style questions. Furthermore, the difference in the number of Type 3 questions highlights the model’s increased ability to recognize incomplete information within a given subcategory.

Clinical Significance-

As AI and LLMs continue to evolve, continuous evaluation of strength and weaknesses are needed to assure that this technology is safely and effectively implemented into clinical practice.