

# How Does a Large Language Model Artificial Intelligence Fare With Basic Science Knowledge?

Zachary C. Lum DO<sup>2</sup>, Dylon P. Collins BS<sup>1</sup>, Lohitha Guntupalli BS<sup>1</sup>, Stanley R. Dennison BS<sup>1</sup>, Soham Choudhury<sup>2</sup>, Augustine M. Saiz MD<sup>2</sup>, Robert Lawrence Randall MD<sup>2</sup>  
<sup>1</sup>Nova Southeastern University, Davie, FL, <sup>2</sup>University of California Davis, Sacramento, CA  
dc2273@mynsu.nova.edu

**Disclosures:** Zachary Lum (N), Dylon Collins (N), Lohitha Guntupalli (N), Stanley Dennison (N), Soham Choudhury (N), Augustine M. Saiz (N), Robert Lawrence Randall (N)

**INTRODUCTION:** As the capabilities of artificial intelligence (AI) continue to advance, it is important to regularly evaluate competency to maintain high standards and preventing potential errors or biases, that could deliver misinformation that could harm patients or spread inaccurate information. A new AI model using large language models (LLM) and non-specific domain areas has gained recent attention in its novel way to process information. We wanted to test its performance to correctly answer basic science questions compared to other subject types and taxonomy question type (recall, interpretation, knowledge application).

**METHODS:** We asked ChatGPT, 3173 questions based on the Orthopaedic In-Training Exam (OITE) and 757 questions from the real OITE. Questions were categorized by subject type, and by taxonomy type. These questions were then entered into the AI chatbot and score was recorded. Multivariate logistic regression analysis was performed comparing basic science questions with other question types, and based upon taxonomy.

**RESULTS SECTION:** After exclusions, ChatGPT answered 960/1871 (51%) of total questions correctly and 254/373 (68%) of basic science questions correctly, which was the highest performing subject type. Basic science exhibited better performance than all subject types except Pathology (p=0.559). Specifically, it performed better than Knee & Sports Medicine (p=0.006), Reconstruction (p<0.001), Spine (p<0.001), Anatomy (p<0.001), Shoulder & Elbow (p<0.001), Hand (p<0.001) and Trauma (p<0.001). When evaluating sub-group taxonomy analysis, univariate logistic regression demonstrated the AI's lower performance in taxonomy type 3 compared to type 1 (50% vs 41%, p=0.049).

**DISCUSSION:** This AI LLM may be most effective in answering orthopaedic questions related to basic science. Furthermore, the study's taxonomy analysis highlights the importance of considering the question structure when evaluating AI performance. Ultimately, as AI continues to evolve and advance, it will be important to consider its limitations and potential biases to ensure its responsible and ethical use.

**SIGNIFICANCE/CLINICAL RELEVANCE:** With the continued growth and integration of artificial intelligence into a variety of tasks, this study reveals that artificial intelligence most effectively demonstrates competency and proficiency with orthopaedic surgeon board examination questions that are specifically focused on basic sciences, especially when compared to performance involving alternative question types.

## IMAGES AND TABLES:

**Table 1:** All questions (OITE + Orthobullets) answered by ChatGPT based upon subject type. BS: basic science, TR: trauma, KS: knee & sports medicine, SP: spine, RE: hip & knee reconstruction, PE: pediatrics, PA: pathology/oncology, SE: shoulder & elbow, HA: hand surgery, FA: foot & ankle, AN: anatomy. The estimated margin means with 95% confidence intervals are displayed.

	Total Questions	Correct	% Correct	Estimated Margin Means (95% CI)
BS	373	254	0.68096515	0.68 [0.33-0.73]
TR	194	81	0.41752577	0.42 [0.35-0.49]
KS	213	121	0.56807512	0.57 [0.50-0.63]
SP	141	77	0.54609929	0.48 [0.39-0.56]
RE	194	77	0.39690722	0.40 [0.33-0.47]
PE	178	94	0.52808989	0.53 [0.46-0.60]
PA	103	67	0.65048544	0.65 [0.55-0.74]
SE	177	53	0.29943503	0.44 [0.37-0.52]
HA	144	53	0.36805556	0.37 [0.29-0.45]
FA	108	53	0.49074074	0.49 [0.40-0.58]
AN	46	15	0.32608696	0.326 [0.21-0.47]
Total for Type	1871	960	0.5130946	

**Figure 1:** Binomial logistic regression based upon 1871 questions, comparing each sub-specialty question type. Orthopaedic basic science comparisons below, with estimated margin means tables and chart listed.

### Binomial Logistic Regression

Model Fit Measures				Overall Model Test		
Model	Deviance	AIC	$R^2_{MCF}$	$\chi^2$	df	p
1	2497	2519	0.0367	95.0	10	<.001

### Omnibus Likelihood Ratio Tests

Predictor	χ <sup>2</sup>	df	p
Question Type	95.0	10	<.001

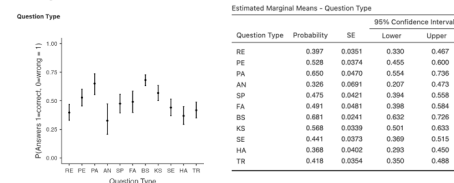
[1]

### Model Coefficients - Answers 1correct, Owingrong

Predictor	95% Confidence Interval				SE	Z	p	Odds ratio	95% Confidence Interval			
	Estimate	Lower	Upper						Lower	Upper		
Intercept	0.758	0.540	0.976		0.111	6.825	<.001		1.717	2.854		
Question Type:												
PE - BS	-0.646	-1.012	-0.280		0.187	-3.457	<.001	0.524	0.364	0.756		
RE - BS	-1.177	-1.537	-0.818		0.184	-6.393	<.001	0.308	0.215	0.442		
PA - BS	-0.137	-0.597	0.323		0.235	-0.584	0.559	0.872	0.551	1.381		
AN - BS	-1.484	-2.138	-0.830		0.334	-4.449	<.001	0.227	0.118	0.436		
SP - BS	-0.858	-1.253	-0.462		0.202	-4.247	<.001	0.424	0.286	0.630		
FA - BS	-0.795	-1.231	-0.360		0.222	-3.578	<.001	0.451	0.292	0.698		
KS - BS	-0.484	-0.832	-0.136		0.177	-2.729	0.008	0.616	0.435	0.872		
SE - BS	-0.997	-1.385	-0.609		0.188	-5.307	<.001	0.369	0.255	0.533		
HA - BS	-1.299	-1.701	-0.898		0.205	-6.323	<.001	0.273	0.182	0.408		
TR - BS	-1.091	-1.450	-0.732		0.183	-5.958	<.001	0.336	0.235	0.481		

Note: Estimates represent the log odds of "Answers 1correct, Owingrong = 1" vs. "Answers 1correct, Owingrong = 0"

### Estimated Marginal Means



**Figure 2:** Binomial logistic regression based upon 724 taxonomy questions. ChatGPT exhibited a higher likelihood of correctly answering a recognition and recall question (Taxonomy 1) versus application of knowledge question (Taxonomy 3) (p=0.049, 1.466 OR [1.002-2.144]). There were no difference between interpretation questions (Taxonomy 2) and recognition and recall questions (p=0.350).

### Binomial Logistic Regression

Model Fit Measures		Overall Model Test			
Model	Deviance	AIC	R <sup>2</sup> <sub>MLR</sub>	χ <sup>2</sup>	p
1	898	1004	0.0040		

### Omnibus Likelihood Ratio Tests

Predictor	χ <sup>2</sup>	df	p
TAXONOMY	4.01	2	0.135

[1]

### Model Coefficients - ANSWERS

Predictor	95% Confidence Interval				SE	Z	p	Odds ratio	95% Confidence Interval			
	Estimate	Lower	Upper						Lower	Upper		
Intercept	-1.21e-15	-0.163	0.18277		0.0933	-1.40e-14	1.000	1.000	0.833	1.201		
TAXONOMY:												
3 - 1	-0.382	-0.763	-0.00184		0.1942	-1.968	0.049	0.682	0.468	0.998		
2 - 1	-0.149	-0.550	0.2524		0.2048	-0.728	0.467	0.882	0.577	1.287		

Note: Estimates represent the log odds of "ANSWERS = 1" vs. "ANSWERS = 0"

### Estimated Marginal Means

