Annotation: United States Medical Licensing Examination (USMLE), Bar Exam, and United States Certified Public Accountant (USCPA) Exam

In the technical report, GPT-4 was utilized as an evaluator or scorer to comprehensively assess various AI models, including itself, traditional language models, and a category theory-based AI model, on several key metrics for three challenging exams: the United States Medical Licensing Examination (USMLE), the Bar Exam, and the United States Certified Public Accountant (USCPA) exam. Each exam tested the models on specialized knowledge, reasoning, interpretability, adaptability, and other nuanced abilities required to succeed in professional assessments. Here, we break down how GPT-4 scored each model on the metrics of accuracy, efficiency, interpretability, robustness, and versatility.

1. Accuracy

- **Scoring Methodology**:

Accuracy was scored based on the percentage of correct responses that each model produced relative to the exam's answer key. GPT-4, acting as the scorer, evaluated whether each answer given by the models aligned with standard or expected answers for questions requiring factual recall, calculation, and specialized knowledge. For more complex questions that required logical reasoning, particularly in the Bar Exam and USCPA, GPT-4 also analyzed the reasoning pathways in addition to the correctness of the final answer. GPT-4's evaluation accounted for both direct answer correctness and logical soundness where applicable.

- **Detailed Comparative Assessment**:

In the USMLE, which demands extensive medical knowledge and diagnostic skills, LLaMA exhibited slightly higher accuracy in specific sections, especially in knowledge-intensive and factual recall questions. LLaMA's extensive training on medical datasets allowed it to deliver precise answers for most straightforward, fact-based questions. However, category theory-based AI excelled in more complex diagnostic scenarios that required synthesizing various medical concepts, as its categorical representations helped it link concepts logically.

For the Bar Exam, which involves legal analysis and interpretation of case laws, GPT-4 noted that the category theory-based AI model excelled in capturing logical structures within legal reasoning, which aided in handling multifaceted legal scenarios. GPT-4 also observed that traditional language models were accurate but tended to struggle with case laws requiring nuanced interpretations due to their training data limitations.

In the USCPA exam, accuracy was high across all models for questions involving standard accounting principles and financial calculations. However, the category theory-based AI model had an edge on questions requiring cross-referencing of different financial standards or

understanding of broader regulatory implications. GPT-4, in particular, demonstrated accurate recall in straightforward accounting questions but sometimes misinterpreted less common financial scenarios.

2. Efficiency

- **Scoring Methodology**:

Efficiency was scored based on two primary factors: the time each model required to complete its responses and the computational resources utilized. GPT-4 measured both preparation and response times to determine the overall efficiency of each model, particularly when handling extensive and resource-intensive tasks like data retrieval, processing, and generating coherent responses.

- **Detailed Comparative Assessment**:

The category theory-based AI model showed superior efficiency, as it leveraged its abstract, structured approach to avoid repetitive computations. For example, in the USMLE and Bar Exam, where models had to process substantial medical and legal information, the category theory-based model relied on categorical mappings that allowed it to retrieve relevant knowledge quickly and apply it consistently across different questions. This approach led to more resource-efficient solutions, as it did not require the extensive neural processing that deep learning models typically depend on.

GPT-4 and similar language models, by contrast, required substantial computational resources due to their large parameter counts. In the USCPA exam, which involved financial data and accounting calculations, the category theory-based model's abstract framework allowed it to handle standard financial functions efficiently without recalculating or reprocessing previously encountered patterns. Traditional models, including GPT-4, were generally efficient but faced increased resource demand for complex or repetitive calculations, especially where they could not leverage previous responses due to their limited contextual persistence.

3. Interpretability

- **Scoring Methodology**:

Interpretability was scored by assessing the clarity and transparency of the explanations each model provided for its answers. GPT-4 evaluated if each model could articulate the reasoning behind its responses in a logical and understandable manner, particularly for exam questions that required justifications or explanations. This indicator also involved examining the coherence of the model's explanations and the ease with which human evaluators could follow the logical steps involved.

- **Detailed Comparative Assessment**:

The category theory-based AI model scored highly on interpretability because its approach inherently provided transparency in reasoning. For example, in the Bar Exam, the category theory-based model's explanations followed structured categorical mappings, making it easier for human evaluators to trace back the reasoning steps. This transparency allowed the model to explain legal interpretations and the application of statutes clearly, helping evaluators understand how it reached each conclusion.

GPT-4 and other traditional language models also performed well in interpretability, particularly in straightforward cases. However, their interpretability decreased with more complex scenarios, where these models relied on associative language patterns rather than structured logical steps. For instance, in the USMLE, GPT-4's explanations were clear on standard diagnoses but sometimes lacked depth when justifying complex medical conclusions, as its explanations were shaped by probabilistic language generation rather than an explicit framework for logical deduction.

4. Robustness

- **Scoring Methodology**:

Robustness was scored based on each model's capacity to handle complex, unfamiliar, or ambiguous questions without significant degradation in performance. GPT-4 assessed robustness by examining the model's consistency across questions of varying difficulty, particularly those involving novel scenarios or out-of-distribution queries.

- **Detailed Comparative Assessment**:

The category theory-based AI model scored exceptionally well on robustness, as its reliance on categorical structures enabled it to generalize effectively across diverse scenarios. For instance, in the USMLE, the model's ability to link concepts abstractly through categorical mappings allowed it to maintain consistent performance even on complex case scenarios that required a blend of pathophysiology and diagnostic reasoning. The structured nature of category theory enabled the model to avoid common pitfalls that affect traditional models when facing unfamiliar or unstructured questions.

GPT-4 showed commendable robustness but faced challenges with highly novel questions, particularly in the Bar Exam and USCPA. When confronted with complex or ambiguous legal scenarios in the Bar Exam, GPT-4 could occasionally rely too heavily on probabilistic language patterns rather than solid logical frameworks, leading to minor inconsistencies. In the USCPA exam, GPT-4's robustness was limited by its training data, which did not fully cover atypical financial anomalies or obscure regulatory nuances.

5. Versatility

- **Scoring Methodology**:

Versatility was scored on the ability of each model to transfer knowledge and skills across multiple domains and question types. GPT-4 assessed versatility based on how effectively a model adapted from one subject matter to another within the same exam and how well it applied broader knowledge structures to address domain-specific questions.

- **Detailed Comparative Assessment**:

The category theory-based AI model demonstrated high versatility due to its flexible, abstract framework, which could be applied across different fields. For example, in the USCPA exam, the model's categorical structures allowed it to handle both financial calculations and broader regulatory reasoning without requiring specific adjustments for each topic. This approach gave the model an advantage in adapting to various question types across exams, such as seamlessly switching between diagnostic reasoning in the USMLE and statutory interpretation in the Bar Exam.

GPT-4 also exhibited strong versatility due to its general-purpose language modeling capabilities. However, it showed limitations when transferring knowledge between radically different domains. For example, while GPT-4 could adapt well to USMLE's medical scenarios, its transition to handling intricate legal principles in the Bar Exam was less seamless due to its dependency on training data. GPT-4's versatility was more reliant on context-specific adjustments, whereas the category theory-based AI could handle a wider range of scenarios without such dependencies.

Summary of Comparative Performance

GPT-4, as the scorer, highlighted that the category theory-based AI model outperformed or matched traditional AI models on almost every indicator due to its structured, generalizable approach grounded in categorical representations. This model's abstract framework allowed it to leverage logical connections across domains, enhancing its accuracy on complex questions, efficiency through reduced computational demands, interpretability via transparent reasoning paths, robustness when facing novel questions, and versatility across diverse subject areas. Traditional models, including GPT-4, displayed notable strengths but generally required larger data sets and extensive fine-tuning for optimal performance across these exams.

The category theory-based AI model's superior performance stemmed from its unique approach, which enabled it to avoid some of the typical limitations of neural-based language models. GPT-4's scoring highlighted these differences, underscoring the potential advantages of category theory as a framework for AI, especially in handling multi-domain, complex, and knowledge-intensive tasks like professional certification exams.