Report of benchmarking EMIR regulation data

1, EMIR on EMIR Regulation-related abbreviations

In examining the performance of various language models on EMIR Regulation-related abbreviations, there are noticeable disparities that highlight underlying weaknesses. GPT-4, while leading with a score of 0.7, and GPT-4 Optimized, achieving 0.78, reveal that even the most advanced models do not reach the high accuracy levels often required for precise regulatory compliance tasks. Notably, GPT-3.5 scores only 0.53, indicating significant limitations in handling specialized terminology. Similarly, Mistral and Llama3, scoring 0.66 and 0.74 respectively, although better than GPT-3.5, still show room for improvement in precision and understanding complex regulatory contexts.

These results demonstrate that current language models, despite their advancements, struggle with detailed and specific regulatory language, which is critical for applications in compliance and legal advisories in finance. The performance gaps, particularly in models not optimized for financial domains, suggest that there remains substantial scope for enhancement in accuracy and domain-specific tuning. These findings point to a clear need for developing more refined models that can better grasp the intricacies of financial regulations to meet the stringent requirements of the sector.