# Report of benchmarking EMIR regulation data

## 1 Tasks and Results

To benchmark EMIR regulation data, our team tested 115 EMIR Regulation-related abbreviations and 50 definitions using various large language models. Here are the results:

| | GPT4 | GPT4o | GPT-3.5 | Mistral | Llama3 |
|---|---|---|---|---|---|
| abbreviations | 0.7 | **0.78** | 0.53 | 0.66 | 0.78 |
| definitions | 0.9 | **0.96** | 0.32 | 0.66 | 0.54 |

Through these tests, we identified some weaknesses in the models.

## 2 Evaluations

### *2.1 EMIR Regulation-related abbreviations*

In examining the performance of various language models on 115 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.

### *2.2 EMIR Regulation-related definitions*

The testing of various language models on 50 EMIR Regulation definitions provides critical insights into their capabilities and exposes some weaknesses. GPT-4 and GPT-4 Optimized lead with impressive scores of 0.9 and 0.96 respectively, indicating a strong ability to comprehend and reproduce complex regulatory definitions. However, the lower performance of other models like GPT-3.5, scoring only 0.32, along with Mistral at 0.66 and Llama3 at 0.54, underscores a broad range of issues in dealing with specialized financial content.

These results highlight a pronounced disparity in model effectiveness, particularly in models not specifically optimized for financial contexts. The poor performance of GPT-3.5 suggests substantial difficulties in adapting to domain-specific language without further training or refinement. Similarly, Mistral and Llama3, though performing better than GPT-3.5, still fall short of the high accuracy needed for reliable application in financial regulation and compliance sectors.

This benchmarking reveals that while some models have adapted well to the complexities of financial regulations, there remains a clear gap in the abilities of general models to handle specialized content accurately. It points to an urgent need for further development and refinement, especially in enhancing the domain-specific performance of models like GPT-3.5, Mistral, and Llama3. This will be crucial for their deployment in sensitive and precision-critical areas such as financial compliance and legal advisory.

# 3 Conclusion

The comprehensive benchmarking of various language models on EMIR Regulation-related abbreviations and definitions reveals both their potential and limitations. While models like GPT-4 and GPT-4 Optimized show promising results, especially in handling complex definitions, there is a consistent theme of underperformance in less optimized models such as GPT-3.5, Mistral, and Llama3 when it comes to specialized financial language.

This benchmarking exercise underscores the necessity for continuous improvement and specialization of language models to meet the demanding accuracy requirements of financial regulations. Future efforts should focus on enhancing model training with domain-specific datasets and employing techniques that might include fine-tuning with targeted financial texts or integrating advanced contextual understanding capabilities. The goal should be to bridge the performance gap and ensure that these tools can reliably support critical tasks in compliance and legal frameworks within the financial sector.

By addressing these weaknesses and refining the models, we can better harness AI capabilities to bolster compliance frameworks, reduce the risk of regulatory breaches, and support the financial industry’s need for precise and dependable regulatory interpretations.