**Dos modelos LLM: Mixtral 8x7B y Llama 2 70B**

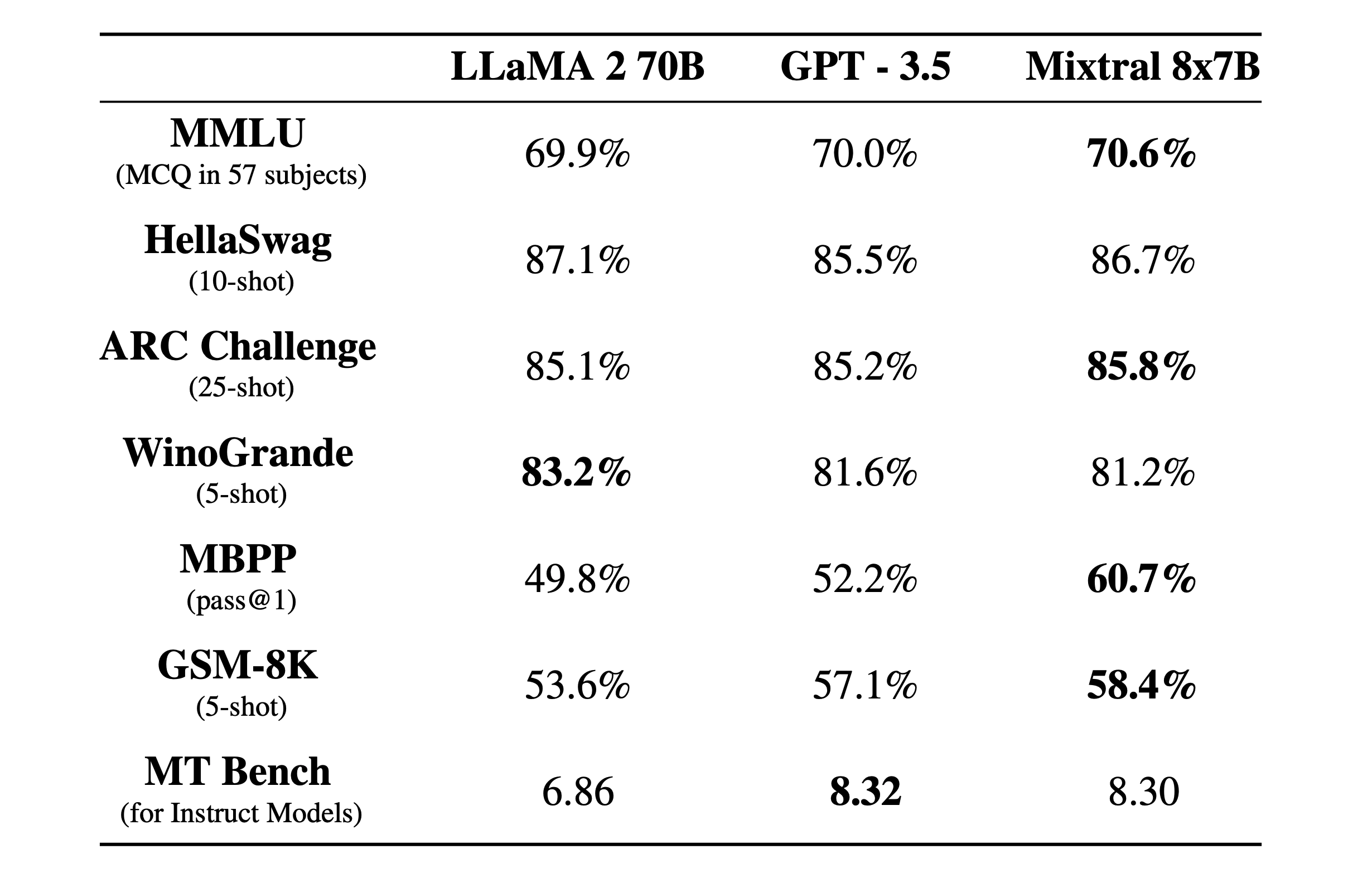
Mixtral 8x7B y Llama 2 70B son modelos de lenguaje grandes (LLM, por sus siglas en inglés) diseñados para diversas tareas de procesamiento del lenguaje natural. Aunque ambos son potentes, tienen diferencias clave en términos de arquitectura, rendimiento y aplicabilidad.

* Mixtral 8x7B es un modelo descrito como una mezcla dispersa de expertos (SMoE, por sus siglas en inglés) que logra resultados de vanguardia entre los modelos abiertos, igualando o superando el rendimiento de modelos como Llama 2 70B y GPT-3.5 Base en evaluaciones. Sus fortalezas específicas incluyen la generación de lenguaje sobre contextos largos (hasta 32,000 tokens), generación de código, y lograr las mejores puntuaciones en seguimiento de instrucciones entre los modelos abiertos. Además, es significativamente más rápido en la inferencia, siendo seis veces más rápido que Llama 2 70B, y está disponible bajo una licencia Apache 2.0, lo que facilita su acceso y uso para proyectos​​.
* Llama 2 70B, por otro lado, es uno de los modelos de la familia Llama 2, que viene en diferentes tamaños de parámetros, siendo 70B uno de los más grandes. Llama 2 utiliza una arquitectura de modelo de lenguaje autoregresiva optimizada, diseñada para una amplia gama de tareas de procesamiento del lenguaje. Aunque es un modelo poderoso, no parece ofrecer la misma velocidad de inferencia que Mixtral 8x7B​​​​.

**Cuadro Comparativo**

|  |  |  |
| --- | --- | --- |
| **Característica** | **Mixtral 8x7B** | **Llama 2 70B** |
| Eficiencia | 6 veces más rápida en inferencia. | Más lenta en comparación, pero robusta en rendimiento. |
| Manejo de Contextos | Puede manejar hasta 32k tokens, ideal para contextos largos. | Capaz en contextos largos, pero posiblemente menos eficiente que Mixtral. |
| Generación de Código | Destacado en generación de código. | Fuerte, pero Mixtral podría tener ventajas específicas. |
| Licencia | Apache 2.0, más abierta para proyectos variados. | Depende del uso y distribución, puede tener restricciones. |
| Arquitectura | Mezcla dispersa de expertos (SMoE), usa una fracción de sus 46.7B parámetros por token. | Arquitectura autoregresiva optimizada para procesamiento de lenguaje. |
| Idiomas Soportados | Inglés, francés, italiano, alemán, español. | Amplia gama de idiomas soportados, dependiendo de la versión y configuración. |
| Optimización | Altamente optimizado para inferencia rápida y manejo de contextos largos. | Optimizado para una amplia gama de tareas de procesamiento del lenguaje. |

Se presenta un análisis comparativo del rendimiento en distintos benchmarks entre la familia GPT-3.5, Llama 2, y Mixtral. De este análisis, Mixtral demuestra una mayor eficiencia en comparación con GPT-3.5 y Llama 2.



Para más información visite: <https://mistral.ai/news/mixtral-of-experts/>

