Comparing Two State-of-the-Art LLMs: Qwen 2.5 7B Instruct vs Mistral 7B Instruct

By: Daniel Mehta

Objectives:

- Understand the design goals and use cases of Qwen 2.5 and Mistral 7B Instruct
- Compare their training approaches, architectures, and data sources
- Evaluate each model's strengths, limitations, and real-world performance
- Highlight key differences in speed, scalability, and accuracy
- Develop insight into how model design affects downstream capabilities

01

What is the main goal or purpose of the model?

Model Purpose: Qwen 2.5 7B Instruct

- Part of the Qwen 2.5 family, open-source LLMs designed for broad general-purpose use
- Optimized for **instruction following, long-context reasoning, and multilingual tasks**
- Fine-tuned to perform well in **chat-style dialogue** and **structured data analysis**
- Excels in **code, math**, and tasks requiring extended coherence
- Post-training enhances alignment with human preferences

Model Purpose: Mistral 7B Instruct

- Instruction-tuned version of **Mistral 7B**, focused on **fast and efficient** general-purpose reasoning
- Designed as a **lightweight, open-source alternative** to larger LLMs
- Intended to demonstrate the **ease of fine-tuning** the base Mistral model for instruction following
- No built-in moderation, meant for **research and developer experimentation**
- Targeted at use cases where performance, speed, and local deployment matter

02

How was the model trained, and what data was used?

Training & Data: Qwen 2.5 7B Instruct

- **Pretrained on 18 trillion tokens**, up from 7T in the previous version
  - Covers common sense, expert knowledge, and reasoning
- Uses **high-quality curated datasets**, likely including web data, code, and multilingual corpora
- **Post-training** includes instruction fine-tuning and alignment to improve:
  - Long-text coherence
  - Structured data understanding
  - Human preference alignment
- Offers both **base and instruct-tuned variants**; quantized versions also available
- Larger proprietary MoE versions (Turbo and Plus) used in Alibaba Cloud Studio

Training & Data: Mistral 7B Instruct

- Built on **Mistral 7B**, a dense 7B-parameter model **optimized for performance and efficiency**
- Uses advanced architecture techniques:
  - **Grouped-query attention (GQA)** for faster inference
  - **Sliding window attention (SWA)** for handling long sequences efficiently
- Trained on a **diverse, curated corpus** (exact datasets not disclosed)
- Instruction-tuned to follow human prompts, resulting in strong general-purpose dialogue performance

03

What are the main strengths and weaknesses of each model?

Strengths & Weaknesses: Qwen 2.5 7B Instruct

Aspect	Strengths	Weaknesses
Knowledge & Reasoning	Deep training on massive token corpus, provides strong reasoning	Lacks live/world event updates
Long-context Handling	Exceptional for long docs, structured generation	
Multilingual Support	Fluent across 29+ languages	
Code & Math	High benchmark scores in programming/math tasks	Needs refinement for complex coding quality
Instruction-Following	Strong alignment via SFT + RLHF	Alignment sensitivity may lead to censorship or bias
Creativity & Conversation	Solid but less imaginative than peers (e.g., Llama, Claude)	
Safety	Generally aligned but susceptible in VL setups	Prompt injection/jailbreak risk in multimodal variants

Strengths & Weaknesses: Mistral 7B Instruct

Aspect	Strengths	Weakness
Benchmarks	Outperforms Llama 13B/34B across tasks	Falls short on in-depth, multi-step reasoning
Efficiency	Fast inference (GQA/SWA), great for edge/home	Lower max context length (~4K tokens)
Accessibility	Fully open-source (Apache 2.0)	No built-in safety guardrails
Real-world use	Highly praised for real-time use and deployment efficiency	Susceptible to hallucination and prompt injection
Language Support	Solid English performance	Less reliable for multilingual use

04

Model Comparison

Performance Comparison: Qwen 2.5 7B Instruct

Speed (vLLM, 1 GPU)

Input Length	BF16 Speed (tokens/s)	GPTQ Int4 Speed (tokens/s)
1	84.3	154.1
6144	80.7	142.0
14336	77.7	129.4
30720	70.3	108.3
63488	50.9	68.0
129024	28.9	26.4

Accuracy

- Outperformed **GPT-4o, GPT-4, and Claude** in a 2024 medical exam benchmark (CNNLE)
- Scored **88.9%** in a 2024 benchmark on China's national medical licensing exam, the highest among 7 major LLMs
- Demonstrated **strong clinical reasoning,** especially in practical and case-based questions

Scalability

- Supports up to **128K tokens**, among the **longest context windows** of any open model
- Available in multiple sizes (0.5B to 72B), with **MoE variants** for cloud-scale deployments
- Scales well on **GPU clusters** using FlashAttention + vLLM backends
- Quantized variants (GPTQ, AWQ) run efficiently on **consumer GPUs**

Performance Comparison: Qwen 2.5 7B Instruct

Scalability

- Fixed **8K context window** using **Sliding Window Attention (SWA)**
- **Grouped-Query Attention (GQA)** allows faster inference with reduced memory load
- Trained with **byte-fallback BPE tokenizer** so it handles out-of-vocab characters efficiently
- Focused on striking a balance between **cost and performance** for smaller deployments

Accuracy

- Outperforms **LLaMA 2 13B** and **LLaMA 1 34B** on reasoning, math, and code generation tasks
- Instruction-tuned version exceeds **LLaMA 2 Chat 13B** in both human and automated benchmarks
- Demonstrates strong performance despite smaller size (7B)

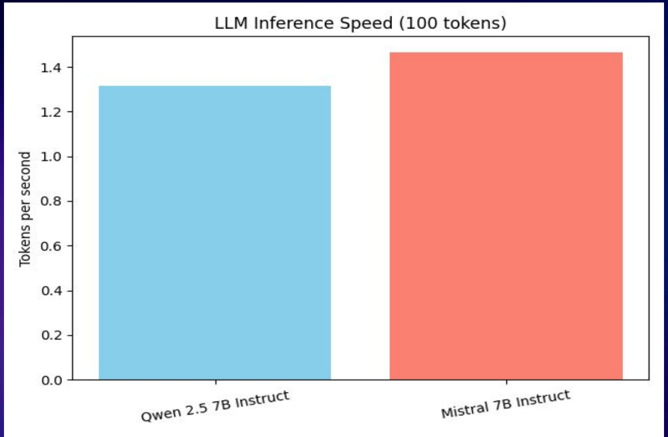
Speed (Inference Engines on A100 GPU)

Inference Engine	Model	Num of prompts	Max token per prompt	Total input tokens	Total Output tokens	Input Token Throughput (Tokens/Sec)	Output Token Throughput (Tokens/Sec)	Execution time(sec)
BUD	mistralai/Mistral-	100	128	**27270**	**12800**	**5584.06**	**2621.05**	**4.88**
vLLM	mistralai/Mistral-	100	128	26967	12800	3826.98	1816.49	7.05
TGI	mistralai/Mistral-	100	128	26967	12750	3898.79	1843.35	6.91

- **Bud Runtime** delivers best performance, ideal for production-scale deployments

- All engines tested with 100 prompts, 128 input tokens, 128 output tokens

Real-World Inference Speed: Qwen 2.5 7B vs Mistral 7B



Test Setup

- **GPU**: RTX 4060 (8GB)
- FP16 Precision (no quantization)
- **Prompt**: "Explain quantum entanglement in simple terms"
- Transformers v4.46

Results

Model	Time (s)	Tokens /sec
Qwen 2.5 7B Instruct	83.72	1.31
Mistral 7B Instruct	75.83	1.46

Main Takeaways

- Mistral ran ~11.5% **faster** in **tokens/sec** than Qwen in local inference
- Qwen may have slightly longer latency due to tokenizer and alignment overhead

References

- Benchmarking Mistral 7B Inference performance on GPUs. Bud. (2024, November 25). https://bud.studio/content/case-study/benchmarking-mistral-7b-inference-performance-on-gpus/
- Das, A. (2023, September 28). Mistral 7B beats llama 2 13B on all benchmarks. DEV Community. https://dev.to/ananddas/mistral-7b-beats-llama-2-13b-on-all-benchmarks-55j2
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023, October 10). Mistral 7B. arXiv.org. https://arxiv.org/abs/2310.06825
- Mistral 7B. Mistral AI. (2023, September 27). https://mistral.ai/news/announcing-mistral-7b
- Mistral. HuggingFace. (n.d.). https://huggingface.co/docs/transformers/en/model_doc/mistral
- Qwen2.5 speed benchmark. Qwen. (n.d.). https://qwen.readthedocs.io/en/v2.5/benchmark/speed_benchmark.html
- Qwen2.5-LLM: Extending the boundary of llms. Qwen. (2024, September 18). https://qwenlm.github.io/blog/qwen2.5-llm/
- Vashisth, V. (2025, February 4). Codestral 25.01 vs Qwen2.5-coder-32B-Instruct: Who codes better?. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2025/02/codestral-25-01-vs-qwen2-5-coder-32b-instruct/
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., … Qiu, Z. (2025, January 3). QWEN2.5 technical report. arXiv.org. https://arxiv.org/abs/2412.15115
- Zhu, S., Hu, W., Yang, Z., Yan, J., & Zhang, F. (2025, January 10). Qwen-2.5 outperforms other large language models in the Chinese National Nursing Licensing Examination: Retrospective cross-sectional comparative study. JMIR medical informatics. https://pubmed.ncbi.nlm.nih.gov/39793017/