Comparative Analysis of FlanT5base, Bart, Mistral7B, and LLaMA2 AI Models

Asilbek Sag'dullayev

June 29, 2025, 06:12 PM +05

Abstract

This research paper conducts a comparative analysis of four AI modelsFlanT5base, Bart, Mistral7B, and LLaMA2examining their architectures, special strengths, GPU usage, prompt tuning support, speed and efficiency, multilingual support, and use cases. The study aims to evaluate their performance profiles and suitability for diverse natural language processing tasks, revealing distinct advantages and trade-offs based on empirical data and literature review.

1 Introduction

Providing background, this study investigates transformer-based AI models' evolution, focusing on FlanT5base, Bart, Mistral7B, and LLaMA2. Research questions explore which model excels across architecture, special strength, GPU usage, prompt tuning support, speed and efficiency, multilingual support, and use case applicability. The rationale, inspired by neural network parallels, seeks to optimize computational efficiency. The hypothesis suggests decoder-only models (Mistral7B, LLaMA2) outperform encoder-decoder models (FlanT5base, Bart) in speed and efficiency, with the approach involving a structured comparison of the specified features.

2 Methods

Detailing the experimental design, this study analyzes the four AI models using a mixed-method approach. The procedure includes collecting data on architecture (e.g., encoder-decoder vs. decoder-only), special strength, GPU usage, prompt tuning support, speed and efficiency, multilingual support, and use cases from existing documentation. Data manipulation involves normalizing VRAM requirements and inference times, while analysis employs comparative tables and descriptive statistics to highlight feature-specific performance.

3 Results

Presenting narrative findings, the analysis delineates model characteristics: - **Architecture**: FlanT5base and Bart use encoder-decoder Transformers, while Mistral7B and LLaMA2 employ decoder-only Transformers. - **Special Strength**: FlanT5base excels

in instruction-tuned task understanding, Bart in denoising autoencoding, Mistral7B in high performance for its size, and LLaMA2 in general-purpose efficiency. - **GPU Usage**: FlanT5base requires 46GB VRAM, Bart 48GB, Mistral7B 1416GB, and LLaMA2 1216GB for 7B models. - **Prompt Tuning Support**: FlanT5base and LLaMA2 support LoRA and PEFT methods, Bart is limited, and Mistral7B supports LoRA/QLoRA. - **Speed Efficiency**: FlanT5base offers 12s/task, Bart is moderate, Mistral7B is very fast with FlashAttention-2, and LLaMA2 is fast. - **Multilingual Support**: FlanT5base is decent (English-strongest), Bart is mostly English, Mistral7B is strong, and LLaMA2 is good. - **Use Case**: FlanT5base suits summarization and QA, Bart for text reconstruction, Mistral7B for chatbots and code generation, and LLaMA2 for reasoning and coding. Figures and tables summarize these metrics, with preliminary data supporting diverse performance profiles.

4 Discussion

Evaluating the study question, this section assesses whether the hypothesis holds across all features. Supporting evidence shows Mistral7B and LLaMA2s speed and efficiency align with decoder-only advantages, while FlanT5bases prompt tuning support and multilingual capabilities challenge encoder-decoder limitations. Interpreting results, each models special strength drives its use case suitabilitye.g., Barts niche in reconstruction. New knowledge highlights GPU usage trade-offs, with unexpected observations of Mistral7Bs multilingual strength despite high VRAM needs. Future studies could benchmark performance, and the conclusion emphasizes model diversity as a key asset in NLP.

A Appendices

This section includes supplementary tables with detailed comparisons to enhance reproducibility.

A.1 Appendix A Model Architecture Comparison

A.2 Appendix B Hardware and Resource Requirements

A.3 Appendix C Multilingual and Uzbek Language Support

A.4 Appendix D Prompt Tuning Compatibility

A.5 Appendix E Use Case Alignment

Model	Architecture	Notes
FlanT5base	Encoder-decoder Transformer	Uses the T5 familys Seq2Seq design. The encoder processes input, the decoder generates output step by step. Ideal for translation, summarization, and Q&A.
Bart	Encoder-decoder Transformer	Similar to FlanT5 but trained as a denoising autoencoder. Good for reconstructing original text from noisy input.
Mistral7B	Decoder-only Transformer (GPT-style)	Autoregressive model. Generates tokens one-by-one, doesn't encode full input ahead of time. Optimized for chat and long generation.
LLaMA2	Decoder-only Transformer	Autoregressive and optimized for efficient inference and reasoning with smaller parameter sizes (e.g., 7B).

Table 1: Model Architecture Comparison

Model	VRAM Needs	GPU Suitability
FlanT5base	46GB	Runs on mid-tier GPUs like NVIDIA T4. Moderate performance.
Bart	48GB	Runs on T4; slightly heavier due to denoising layers.
Mistral7B	1416GB (810GB with quantization)	Requires A100 or similar; quantization enables Colab T4 usage.
LLaMA2	1216GB	Similar to Mistral; efficient inference, needs high memory.

Table 2: Hardware and Resource Requirements

Model	Multilingual Strength	Uzbek Support
FlanT5base	Decent (English-focused)	Limited can handle basic Uzbek phrases.
Bart	Mostly English	Very limited support, not suitable for Uzbek.
Mistral7B	Strong	Good trained on diverse multilingual corpora, including potential Uzbek exposure.
LLaMA2	Good	Moderate better than Bart or FlanT5, with multilingual training data.

Table 3: Multilingual and Uzbek Language Support

Model	Prompt Tuning Support	Notes
FlanT5base	LoRA, Prefix Tuning (PEFT)	Easily extendable via PEFT libraries for lightweight fine-tuning.
Bart	Limited	Fine-tuning possible, but less compatible with modern PEFT methods.
Mistral7B	LoRA, QLoRA	Efficient and widely used for low-rank adaptation.
LLaMA2	LoRA, QLoRA	Highly adapted for PEFT-based customization.

Table 4: Prompt Tuning Compatibility

Model	Recommended Use Cases
FlanT5base	Summarization, Q&A, translation, classification, reasoning especially in prompt-based tasks.
Bart	Summarization, noisy text generation, grammar correction, sentence reconstruction.
Mistral7B	Chatbots, long-form text, code generation, Q&A, general-purpose text AI.
LLaMA2	Reasoning, coding, document generation, chatbot AI, zero/few-shot tasks.

Table 5: Use Case Alignment