LLM Comparison Report

# 📊 Comparison: Mistral-7B-Instruct vs. LLaMA-3-8B-Instruct

|  |  |  |
| --- | --- | --- |
| Feature | Mistral-7B-Instruct | LLaMA-3-8B-Instruct |
| License | ✅ Apache 2.0 (fully commercial) | ❌ CC-BY-NC 4.0 (non-commercial only) |
| Model Size | 7B | 8B |
| Tokenizer | SentencePiece BPE | Tiktoken (OpenAI-compatible) |
| Context Window | ~32k tokens | 8k tokens |
| Multilingual Support (EN/FR) | ⚠️ English-optimized; French usable with tuning | ⚠️ Mainly English; struggles in French extraction tasks |
| Instruction Following | ✅ Strong and minimal | ⚠️ Verbose; repeats instructions and outputs |
| Clean JSON Extraction | ✅ Excellent on short/medium English CVs | ❌ Repeats question + multiple copies of JSON |
| Prompt Repetition | ❌ Very low when return\_full\_text=False | ❌ Frequent; even with return\_full\_text=False |
| Handling Long CVs | ⚠️ May truncate if prompt not optimized | ❌ Often fails, generates noisy output |
| Token Efficiency | ✅ High (compact, clean output) | ❌ Poor due to duplication of instructions and outputs |
| Speed (Inference) | ✅ Fast (7B, efficient) | ⚠️ Slightly slower, larger model |
| Fine-tuning Ready | ✅ Yes (open weights, permissive license) | ❌ No (license forbids commercial fine-tuning) |

# 🧪 Observations Summary

✅ Mistral-7B-Instruct:

* - Clean JSON extraction from English CVs
* - Slight degradation on long/dense French CVs, but no hallucinations
* - Very efficient and compact outputs

❌ LLaMA-3-8B-Instruct:

* - Repeats prompt and output multiple times
* - Does not obey "only JSON" instruction reliably
* - Not suitable for commercial use or fine-tuning

# 🔧 Recommendation

If you want to build a production-grade, multilingual CV parser: ✅ Use Mistral-7B-Instruct (or Mixtral-8x7B-Instruct if possible)

If you’re prototyping non-commercially: LLaMA-3-8B-Instruct may be acceptable, but with limitations.