# SAPCAD LLM Fine-Tuning – Updated Model Selection Guide

## Objective

Train an LLM that:
- Understands architectural design instructions (English/German)
- Maps instructions to IFC elements with BIM standards compliance
- Generates accurate JSON outputs for add, modify, remove actions for use with IfcOpenShell

## Recommended Models

| Model | Why Choose | Advantages | Disadvantages |
|---|---|---|---|
| LLaMA 3 8B / 13B (Meta) | Strong multilingual understanding, excellent for structured outputs | High accuracy, good English/German support, strong community | Requires high-end GPU (24GB+ VRAM recommended) |
| Mistral 7B | Lightweight open-weights, good at JSON and instructions | Lower resource needs, good for structured data | Slightly weaker in multilingual handling |
| Qwen 7B | Good for command-following and structured JSON output | Balanced performance/size, solid JSON generation | Less optimized for German |
| Phi-3 3.8B | Best for low-resource environments | Small, efficient, fast | Lower accuracy on complex instructions |
| Phi-2 (used for testing in Colab) | Chosen for prototyping on limited hardware (Colab Free) | Loads in 4-bit quantization, works on Colab Free, good for small-scale test | Not ideal for production, less accurate on large/complex instructions |

## Summary of Current Choice

Phi-2 (microsoft/phi-2) was selected for initial LoRA fine-tuning and testing in Colab due to:
- Low VRAM requirement (worked with 4-bit quantization)
- Ability to complete LoRA training on Colab Free
- Successful generation of structured JSON for add/remove/modify actions

## Recommended model for university server

If the university server has access to A100 / H100 / 3090 / 4090 (>=24GB VRAM):
→ LLaMA 3 8B or 13B is recommended because:
- Best performance for multilingual (English/German) instructions
- Higher accuracy in structured JSON generation
- Suitable for complex architectural tasks

If server has medium GPU (16-24GB VRAM):
→ Mistral 7B + LoRA would be a great choice

## Next steps

- Use Phi-2 outputs to validate prompt/response structure
- Prepare university server for full fine-tuning with LLaMA 3 8B or Mistral 7B