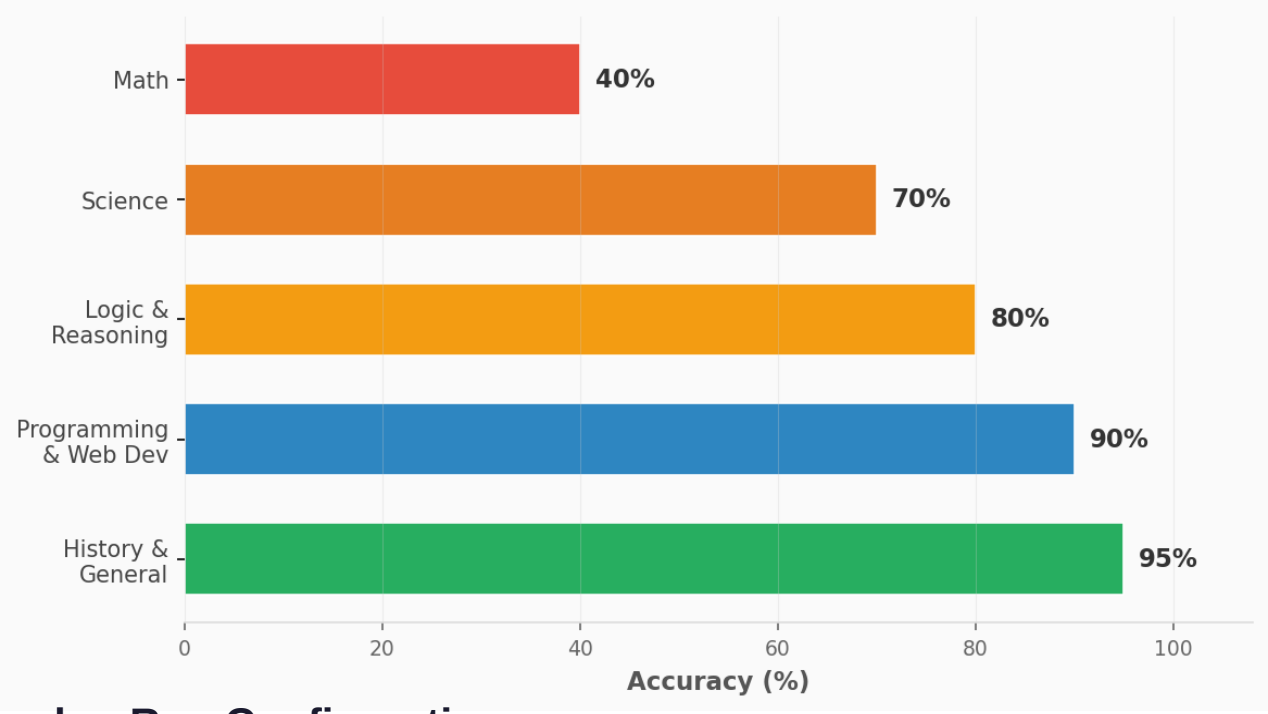
LLM Inference Benchmark Report

Hardware: 4x RTX 3060 (NVMe) | Engine: llama.cpp server-cuda13 | ~51 tok/s

Model: Qwen3.6-35B-A3B-MoE (Q6_K) + Opus-to-Kimi LoRA

Accuracy by Category



Docker Run Configuration

```
sudo docker run --rm -p 8080:8080 \
  -v /media/nvme_storage/model/:/models \
  --gpus all --ulimit memlock=-1:-1 \
  --env CUDA_VISIBLE_DEVICES=0,1,2,3 \
  ghcr.io/ggml-org/llama.cpp:server-cuda13 \
  -m /models/llmfan46_Qwen3.6-35B-A3B...gguf \
  --lora-scaled ...opus-to-kimi-lora.gguf:0.6 \
  --mmproj /models/mmproj-BF16.gguf \
  --host 0.0.0.0 --port 8080 \
  --n-gpu-layers 999 --tensor-split 10,12,12,12 \
  --ctx-size 131072 --batch-size 8192 \
  --ubatch-size 512 --cache-type-k f16 --cache-type-v f16 \
  --flash-attn on --cont-batching --mlock \
  -n 8192 --no-mmap --parallel 1 \
  --chat-template-file /models/chat_template.jinja \
  --temp 0.4 --top-k 20 --top-p 0.9 --min-p 0.05 \
  --xtc-probability 0.5 --xtc-threshold 0.1 \
  --presence-penalty 0.1 --frequency-penalty 0.1 \
  --repeat-penalty 1.1
```