Plans for the knowledge base

MARIETTA BLAU INSTITUTE FOR PARTICLE PHYSICS

Until the end of the year

- Measurements for the various components
- Conversion to a service with chat interface and MCP server
- Draft for publication

Longer term

Depends on interest beyond DELPHI

- Conversion of the pipeline to a service
- From a static LLM wiki to an updateable knowledge base with newly arriving documents

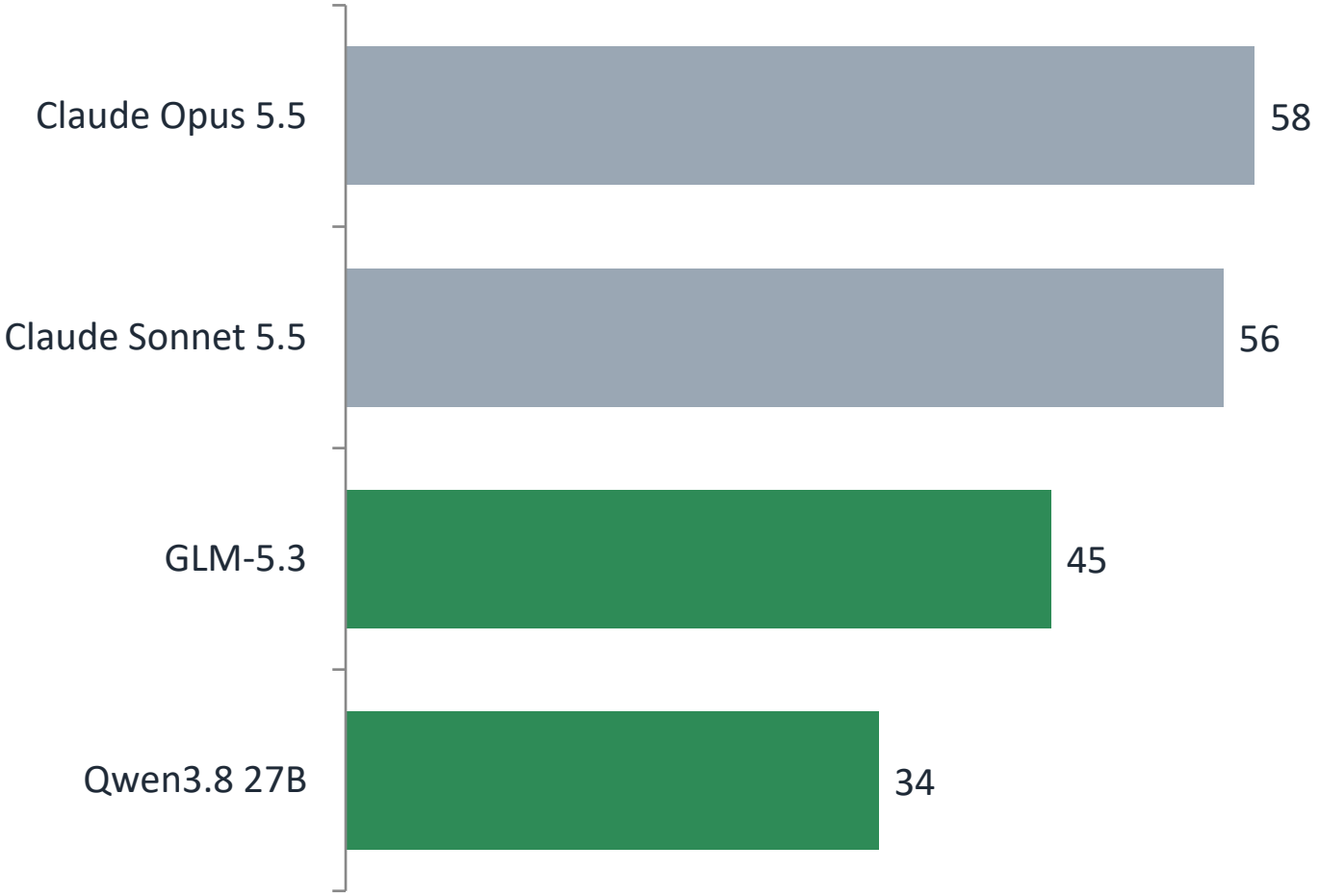
Open-source LLMs and CERN plans

MARIETTA BLAU INSTITUTE FOR PARTICLE PHYSICS

- CERN will not buy Anthropic or other licenses
- CERN will offer open-source models like Qwen3.8 27B and GLM-5.3

This work would not have been possible without these models

Model comparison



Open weights Proprietary

GLM-5.3 is estimated to be several months behind the leading frontier models, more for agentic work

"Open models lag state-of-the-art closed models by 4 months"

Epoch AI, Data Insight, 29 May 2026

Index: Artificial Analysis Intelligence Index v4.3.2 (October 2026), best setting of each model; not comparable with the older v4.1.1 scores.

Hardware required

MARIETTA BLAU INSTITUTE FOR PARTICLE PHYSICS

Model	Precision	Memory	Typical hardware
Qwen3.8 27B	BF16	~54 GB	1× H100 80 GB
GLM-5.3	NVFP4	~420 GB	4–8× B200 (CERN on site)
GLM-5.3	FP8	~750 GB	8× H200 / B200 (CERN on Google Cloud)

Approximate, weights only; long contexts need additional memory for the KV cache.