

A Malay AI You Can Run Yourself

Sam Wong · sam.wong@oaica.com · Oaica (oaica.com) · 2 October 2026

Oaica develops Malay language models for organisations considering local AI deployment. Oaica 35B-A3B Malay v1.0 260923 and its safety-tuned variant Oaica 35B-A3B Malay Safety v1.0 260923 are fine-tuned from the open-weight base Qwen3.6-35B-A3B. The same weights are not tied to one accelerator class: our `.oqm` serving engine can pin the model’s mixture-of-experts layers in host CPU RAM, filling the available GPU memory automatically and off-loading the remainder, so the family runs on laptop- and desktop-class GPUs at reduced context as well as on datacenter servers. The weights ship in our own `.oqm` quantised container, and higher-throughput engine configurations are offered for API and cloud inference. The required runtime and context memory depend on the workload. An installation configured without external services or outbound integrations can keep inference on the customer’s premises, including in an air-gapped environment.

What has been measured

- **Malay knowledge.** Historical general and safety variants scored 82.8 and 83.4 on our 500-question MalayMMLU sample with reasoning on. The recorded sampling intervals were about ± 3.3 points; these are not full-set scores or a matched comparison with stock models.
- **Translation.** Historical automatic aggregate scores were 90.8–91.0. These are metric scores, not the percentage of translations judged correct by people.
- **Memory and serving.** A 4-bit format reduced weight memory by about 3.2 times. The paper reports the GPU, runtime, prompt length and concurrency behind the historical serving measurements, which used the previous A100/vLLM stack; our current `.oqm` serving engine targets Blackwell-class hardware (sm_120, RTX PRO 6000).
- **Cross-language check.** The safety variant and its base model were also evaluated on an Indonesian SEA-HELM set, a language they were not tuned for, to check that the measurement issues are not specific to Malay. The figures are in the paper; the check ranks no model.

These historical results have incomplete checkpoint and interval provenance. The paper separates them from recomputed safety runs and publishes a companion evidence document with text-free counts, code and outstanding evidence gaps.

Start with a supervised pilot

Malay drafting, translation and content screening are candidate uses. A pilot needs tests on the organisation’s actual domain, language mix and policy. For screening, measure harmful-content misses as well as false alarms. For drafting and translation, check factuality and meaning. Deployments should route uncertain cases to a person; an escalation workflow is a deployment requirement, not a feature established by these benchmarks.

Oaica 35B-A3B Malay Coder v1.0 260923 and Oaica 35B-A3B Malay Researcher v1.0 260923 are experimental models. Their coding, knowledge and screening results vary by test; neither is validated for unreviewed decisions. Long-context and broader safety evaluation remain incomplete.

What the paper does not establish

The study does not establish that any model is safe, safer than another, regulator-approved or contamination-free. Low toxicity scores do not identify a flaw in the benchmark’s labels. Overlapping score

intervals do not by themselves establish that models are equivalent. Local deployment gives an organisation control over its configuration; it does not replace security, policy or task-specific validation.

Read the case study. Contact info@oaica.com for local-deployment enquiries or research@oaica.com for the evaluation evidence.