

Malay AI Safety Scores Measure Less Than You Think

Oaica Malay models and the evidence for supervised pilots

Sam Wong · sam.wong@oaica.com · Oaica (oaica.com) · 1 October 2026

Oaica develops Malay language models for local deployment. Our case study shows how aggregation and inference settings can change observed safety scores. It does not establish that any model is safe or that one model is safer than another. A companion evidence document provides text-free counts and score-recomputation code, and identifies the provenance still missing from the historical record.

Local deployment

Oaica 35B-A3B Malay v1.0 260923 and Oaica 35B-A3B Malay Safety v1.0 260923 have about 35 billion parameters in total and about 3 billion active per token. They 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 offloading the remainder, so the same weights run from laptop- and desktop-class GPUs at reduced context up to datacenter servers. The weights ship in our own `.oqm` quantised container, and higher-throughput engine configurations are offered for API and cloud inference. Runtime and context memory must also fit the intended workload. A fully local or air-gapped installation can keep inference on the customer’s premises when external services and outbound integrations are disabled.

Oaica 35B-A3B Malay is fine-tuned from Qwen3.6-35B-A3B, the open-weight base; the stock reference run measures that base directly. Final checkpoint hashes and their mapping to the historical results remain to be verified; the product name alone is not an immutable model identifier.

What the evidence shows

- **Historical safety measurements:** 47.2 for the safety variant and 45.6 for the general variant, using our reproduction of the SEA-HELM Malay formula. The historical analysis reported 95% item-sampling intervals of roughly ± 6 points. These are archived results, not a matched comparison with the newer runs below; the original interval reconstruction remains incomplete.
- **Historical MalayMMLU:** 83.4 and 82.8 on our 500-question sample, reasoning on, with reported intervals of about ± 3.3 points. Stock 80.7 was measured under a different full-set first-token protocol; no Malay-knowledge uplift over stock follows from that comparison.
- **Cultural proxy:** 76.8–76.9 on our Malay natural-language-inference task. It does not measure the same quantity as the leaderboard’s cultural competency and does not validate culturally sensitive moderation.
- **Translation:** aggregate automatic scores of 90.8–91.0 for the historical released checkpoints. Separate Chinese-direction scores came from development checkpoints. These are metric scores, not percentages of translations judged correct by people.
- **Benign over-refusal:** a historical range of 1.7–8.5% across variants. This alone says nothing about harmful-content misses, subgroup errors or adversarial robustness.
- **Indonesian cross-language probe:** the safety-tuned checkpoint and its base model were also run on the Indonesian SEA-HELM set, a language neither was tuned for, to test whether the measurement issues above carry across languages. No Indonesian-targeted text entered the fine-tuning mix. The figures are reported in the paper; the probe ranks no model and makes no comparative safety claim.

Recomputed safety runs

The following saved runs were recomputed from verified item files. They use the same four-task balanced-accuracy formula, greedy generation with reasoning off and batch size 8. Each has 1,572 scored items. Model identity follows the local run records; immutable checkpoint verification is pending.

Run and recorded model	Composite	Conditional 95% interval
Saved Run 1 — Oaica 35B-A3B Malay v1.0 260923	44.92	37.88–51.58
Saved Run 2 — Oaica 35B-A3B Malay Coder v1.0 260923 (experimental)	45.91	39.40–51.54
Saved Run 3 — Oaica 35B-A3B Malay Researcher v1.0 260923 (experimental)	45.24	38.60–51.24
Saved Run 6 — Oaica 35B-A3B Malay Safety v1.0 260923	47.31	40.36–53.66

The intervals describe item-sampling uncertainty conditional on the saved outcomes and preserve shared cultural prompts. They exclude training, generation-rerun and selection uncertainty. These figures are computed from verified saved outcomes and do not overwrite the historical runs above. The two runs not listed here are the stock Qwen3.6-35B-A3B and SEA-LION v4.5 27B references, given in the paper and supplement. No paired model-difference test or deployment-safety ranking is claimed. Full definitions, run hashes and the separate stock-reference measurements are in the paper and supplement.

Serving measurements

On one NVIDIA A100-80GB with vLLM 0.30.0, bfloat16 or 4-bit weights, one stream, a 2,864-token prompt and up to 200 output tokens, historical effective prefill throughput was 14,300–15,600 tokens/s and time to first content token about 0.2 seconds. Prefix-cache hits were excluded. These are historical measurements on the previous serving stack, not latency guarantees. Our current .oqm serving engine targets Blackwell-class hardware (sm_120, RTX PRO 6000), with FP8-only kernels and split-KV and shared-quantisation paths that replace the A100 baseline; the A100 figures are retained only as a historical reference point. The 4-bit format reduced weight memory by about 3.2 times; memory savings alone do not establish equivalent model behaviour.

What a pilot must establish

Moderation, drafting and translation are candidate uses for supervised pilots. Each needs validation on the customer’s policy, domain and language mix, including false alarms, harmful-content misses, factuality and human review. Deployments should route uncertain cases to people; this publication does not establish an implemented escalation feature or a completed native-speaker audit. Benchmark contamination checks and independent review remain incomplete.

Our historical toxicity scores were low, but their cause is unresolved. They do not establish a benchmark ceiling or defects in its labels. Overlapping model intervals likewise do not settle whether paired differences are significant. Procurement decisions need the actual deployment evidence; this work claims no government endorsement or regulatory compliance.

Further information

Read the case study and evidence document. Oaica also maintains benchlint, a reporting-hygiene tool; its checks are not model certification.

Enquiries about local deployment or evaluation: info@oaica.com. Research: research@oaica.com.