

Oaica 35B-A3B Malay v1.0 260923 model family notes

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

These notes describe the general and safety variants and two experimental specialists. The family is fine-tuned from the open-weight base Qwen3.6-35B-A3B. The figures are local measurements, with protocol and provenance limits. None establishes comparative deployment safety or readiness for unreviewed decisions. The safety variant and its base model were also run on an Indonesian SEA-HELM set, a language neither was tuned for, to check that the measurement issues hold across languages; those cross-language figures are reported in the paper and rank no model.

1. Candidate uses

Model	Candidate use in a supervised pilot	Evidence still needed
Oaica 35B-A3B Malay v1.0 260923	Malay drafting and assistance on local hardware	Domain factuality, instruction-following and workload validation
Oaica 35B-A3B Malay Safety v1.0 260923	Malay screening with human review	Harmful-content misses, false alarms, subgroup and adversarial tests
Oaica 35B-A3B Malay Coder v1.0 260923 (<i>experimental</i>)	Code generation from Malay prompts	Broader coding tests and review of generated code
Oaica 35B-A3B Malay Researcher v1.0 260923 (<i>experimental</i>)	Malay instruction-following and drafting	Source verification, factuality and long-context evaluation

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. Runtime and context memory still require workload-specific budgeting. Fully local operation requires external services and outbound integrations to be disabled. Checkpoint names in these notes follow the local records; immutable checkpoint-to-product mapping remains pending.

2. Knowledge and coding measurements

MalayMMLU below uses **484 scored questions**, greedy generation with reasoning off and answer-letter extraction. It is a separate protocol from the paper’s historical 500-question, reasoning-on results of 82.8 and 83.4. HumanEval uses 164 coding problems. Figures describe each test separately; percentages across different tests are not interchangeable. Counts are included to make the denominators explicit. No paired significance claim is made.

Recorded model	MalayMMLU correct / 484	HumanEval passed / 164
Oaica 35B-A3B Malay v1.0 260923	392 (81.0%)	67 (40.9%)
Oaica 35B-A3B Malay Safety v1.0 260923	387 (80.0%)	52 (31.7%)
Oaica 35B-A3B Malay Coder v1.0 260923 (experimental)	351 (72.5%)	77 (47.0%)
Oaica 35B-A3B Malay Researcher v1.0 260923 (experimental)	365 (75.4%)	129 (78.7%)
Stock reference	Not interpretable as knowledge under this extraction protocol	126 (76.8%)

These are descriptive point estimates, not population guarantees. The stock reference’s 3–4 extracted correct letters out of 484 are a format/extraction failure; we exclude them from capability comparisons. The paper reports 80.7 for a separate full-set first-token run of that reference, which is also not comparable to the sample figures here. No 80-point Malay capability uplift or recipe superiority is claimed. The safety variant holds knowledge close to the general variant (387 against 392) and its HumanEval figure is lower (52 against 67); both are recorded for completeness and are not offered as coding results. The reason for the HumanEval difference is not established here.

3. Recomputed safety measurements

Each run uses 1,572 scored items and our reproduction of the SEA-HELM Malay balanced-accuracy formula: toxicity plus three cultural subtasks. Greedy generation, reasoning off, batch size 8. The intervals below describe conditional item-sampling uncertainty, retaining shared prompt/response structure; they exclude generation reruns, training and model selection.

Saved 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 unreconciled specialist-panel values are not included here. Saved Runs 4 and 5 are the stock Qwen3.6-35B-A3B and SEA-LION v4.5 27B references, which the evidence document lists alongside Runs 1–3 and 6. The historical safety-variant 47.2 and general-variant 45.6 come from an earlier serving stack and generation protocol and are not the figures in this panel; the safety variant’s recomputed figure here is Run 6. No safety ranking follows from overlapping marginal intervals; a paired comparison would be needed. The evidence document provides the counts, run hashes and score-recomputation code and identifies missing provenance.

4. Specialist limits

On HumanEval Oaica 35B-A3B Malay Coder v1.0 260923 passed 77 of 164 problems, compared with 126 for the stock reference. A separate LiveCodeBench v6 run reported 28.6% on 175 problems; the different benchmark does not itself contradict HumanEval. On a small, easy 16-item Malay-instruction coding set, it passed all 16; this is a feasibility observation, not broad validation. Oaica 35B-A3B Malay Researcher v1.0 260923 passed 129 of 164 HumanEval problems, but that alone does not validate

research ability. Neither specialist has completed the domain, safety and over-refusal evaluation needed for a deployment recommendation. Long-context evaluation remains incomplete.

5. Permitted claims

Use measured values with their protocol, denominator, run and limitations. Describe local deployment as a configuration to validate. Describe moderation and drafting as supervised pilot candidates. Do not claim safety superiority, regulatory approval, verified absence of contamination, a universal capability ceiling or replacement of human reviewers.

Research: research@oaica.com. General enquiries: info@oaica.com. Full paper. benchlint checks reporting hygiene; it does not certify a model.