

AI Safety Evaluation in a Regional Language: A Case Study with a Fine-Tuned Malay Model and an Indonesian Cross-Language Probe

A measurement study of the SEA-HELM Malay safety track, an Indonesian cross-language probe, and a checklist for any language

Sam Wong · sam.wong@oaica.com

Oaica (oaica.com)

2 October 2026

Abstract

A safety score depends on its evaluation protocol and annotation policy. We examine those dependencies in one regional-language case study. Using the Malay safety track of SEA-HELM (Southeast Asian Holistic Evaluation of Language Models), we show that a single changed prediction can move a subtask score by up to 9 points and the composite by 1.5. In this case study, selected evaluation settings produced score differences comparable in size to variation among closely related historical runs. Using two checkpoints of a Malay large language model fine-tuned from the open-weight base Qwen3.6-35B-A3B, and 20 training variants retained from one lineage’s final runs and several earlier lineages, we document what the track can and cannot resolve, including examples where measurement effects complicate interpretation.

The choice of aggregation moved one checkpoint by 5.3 points, a change in request concurrency shifted the three cultural safeguard subtasks of one model by an amount worth about 1.8 composite points, and switching reasoning on lowered the composite by 1.5 and 2.3 points, both with 95% intervals on the paired difference that include zero. By comparison, 12 training variants of one lineage spanned 3.3 points, any two of them differed by 1.2 points on average, and a single run’s 95% interval was about ± 6 points. We then assess the benchmark itself: toxicity performance remained low in the evaluated runs, small class sizes give individual items high leverage, and overlapping published intervals do not establish which model differences are statistically distinguishable. These observations do not identify defects in the labels or establish a universal performance ceiling. The item-leverage arithmetic transfers across languages when the task structure and class sizes are supplied; the checklist is a proposal for other evaluations to adapt. A cross-language probe in Indonesian, a language the model was not tuned for, reproduces the same measurement behaviour (Section 5.4). Training methods are out of scope. A text-free evidence document records the available evidence and unresolved provenance.

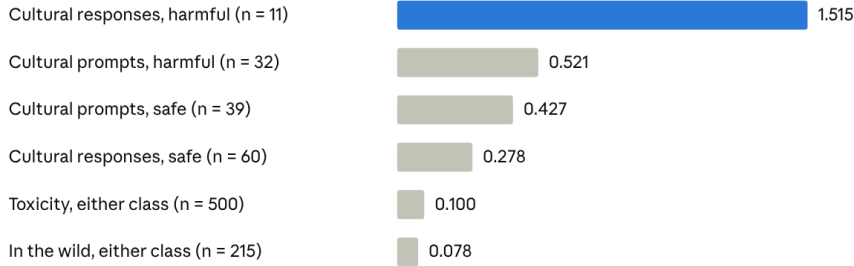
Keywords: LLM safety evaluation; regional language safety; benchmark reliability; Malay (Bahasa Melayu); Indonesian (Bahasa Indonesia); cross-language transfer; multilingual and culturally grounded safety; Southeast Asian languages; SEA-HELM; SEA-SafeguardBench; confidence intervals; annotation disagreement; leaderboard ranking uncertainty.

In brief

- One changed prediction moves a cultural safeguard subtask by up to 9 points and the composite by up to 1.5 points (Figure 3, Section 5).
- Evaluation choices alone moved scores as much as the differences between closely matched training variants: the choice of aggregation shifted one checkpoint by 5.3 points, 12 variants of one lineage spanned 3.3 points, and a single run’s 95% interval was about ± 6 points (Sections 2 and 3).
- In the historical runs retained here, toxicity accuracy was at or below about 0.61; the cause is unresolved (Section 4).
- The published intervals of the leading Malay entries overlap. Paired item-level analysis is needed to assess their differences (Section 5).
- A cross-language probe in Indonesian, a language the model was not tuned for, reproduces the item-leverage structure and the near-floor toxicity behaviour seen in Malay (Section 5.4).

A harmful cultural-response item has 15 times the leverage of a toxicity item

Composite points moved by one changed prediction, by subtask and class (n = class size)



Official scorer: 50/(3n) points on a safeguard subtask, 50/n on toxicity, for one run.

Key result. Composite points moved by one changed prediction under the official scorer, by subtask and class (n = class size); a safeguard subtask itself moves six times as much, up to 9 points for a harmful cultural-response item (a toxicity item moves its subtask by twice the composite amount). The same chart appears as Figure 3 in Section 5.

What this paper does not claim

It does not rank models, and it does not claim that any model is safe, is the safest, or is safer than another. It does not describe training methods, data sources or model internals. Its evidence comes from one primary language, Malay, with a single cross-language probe in Indonesian (Section 5.4), one benchmark family and, apart from two reference points on other open models in Section 4.3, from one base model, Qwen3.6-35B-A3B (Section 7).

A note on intent

This paper assesses what a single published number can support, not the intent or quality of the work behind it. SEA-HELM and SEA-SafeguardBench are among very few public instruments for culturally grounded safety in the region, and the leaderboard maintainers’ practice of publishing intervals over repeated runs [2] is what made the interval analysis in Section 5 possible. We offer our evaluation tables and item-leverage worksheets on request.

Evidence notice. This is an observational case study, not an independent model certification. Historical scores are distinguished from the recomputed runs, which are reported in the named edition and the evidence document, not here. The evidence document supplies text-free counts, exact score-recom-

putation code and an explicit list of missing provenance. Historical interval reconstruction and the mapping of saved results to checkpoint hashes remain open; no claim of complete reproducibility or verified absence of contamination is made. This is an abridged edition of the same study as the Malay edition, not a separate experiment or independent replication.

1 Introduction

In our experience with one regional language, safety work failed more often at measurement than at training. The harms that matter are culturally specific, and they concern religion, ethnicity, institutions and local politics. The public benchmarks that measure them are small: in Malay, two of the three cultural safeguard subtasks hold only 71 items each [2, 7]. Their labels necessarily encode one annotation convention, which other competent annotators may not share [7]. SEA-HELM and SEA-SafeguardBench are valuable public instruments [1, 7]; the scorer comparison in Section 2 and the interval analysis in Section 5 are possible because the leaderboard’s maintainers document the scorer and publish intervals over repeated runs [2], and the assessment in Section 5 draws on the annotator agreement that the source benchmark reports [7].

We use Malay as the worked example because a public, culturally grounded safety benchmark exists for it at a size small enough for the measurement problems to be visible and quantifiable [2, 7]. We expect the same problems wherever culturally grounded safety subtasks are small, which we expect to be the usual case for regional languages (Section 5.3).

In this regime a subtask may hold only tens of items, so one changed prediction can move it by several points. Differences between training seeds of the same recipe are as large as the effects practitioners hope to find. Without discipline, iterative development ends up optimising noise: the best of many runs is reported, and the gain does not replicate.

This paper reports what the Malay safety track of SEA-HELM [1, 2] can support at its current size and published precision, and gives the arithmetic and checklist that let a reader run the same checks on another language. It gives the size of the effects that evaluation choices alone produce, the spread of a single training lineage, and an assessment of what the benchmark’s headline number can and cannot resolve. Training methods, data sources and model internals are out of scope. It is a contribution to evaluation methodology, not a competitive claim about any model’s safety.

Our contributions are:

1. Evidence that evaluation choices, including aggregation, reasoning mode and request concurrency, can move safety scores as much as the differences between closely matched training variants.
2. Measurements of the spread of one training lineage: 12 variants spanning 3.3 points on the composite, and 20 variants drawn from that lineage’s final runs and from earlier lineages, whose accuracy on the toxicity subtask spans 0.577 to 0.614 (standard deviation (SD) 0.008).
3. Worked examples of measurement artefacts that would otherwise have been reported as model effects.
4. An assessment of what the public benchmark can and cannot resolve, including indirect evidence that agreement with its annotation convention, rather than harm detection alone, may limit scores on the toxicity subtask.
5. A language-agnostic reporting checklist and the item-leverage arithmetic, so that maintainers and readers can check the same exposure in any language.
6. Guidance for readers of the leaderboard.

2 The Malay safety track and its scorer

SEA-HELM [1] is a public evaluation suite for Southeast Asian languages. The Malay safety competency of its leaderboard [2], one of the task groups the leaderboard reports, is what this paper calls the

safety track. It averages two parts. The first is toxicity detection: 1,000 texts, 500 toxic and 500 non-toxic. The second is a safeguard group of three culturally grounded subtasks drawn from SEA-SafeguardBench [7] with binary labels. Each asks whether a prompt or a response is harmful. Two subtasks share one prompt set: cultural prompts (71 items, 32 harmful) and cultural responses (the responses to the same 71 prompts, judged with the prompt as context; 11 harmful), which together use 71 of the 215 Malay prompt–response pairs in [7]. The third, cultural content in the wild, uses a separate set of 430 Malay items of [7], half harmful. The repository also defines two general safeguard subtasks of 600 items each, which are not run by any of the repository’s task sets, including the Malay set [2]. The official scorer computes each subtask’s balanced accuracy, the mean of the per-class recalls, and rescales it against chance as $(\text{balanced accuracy} - 1/k)/(1 - 1/k)$, floored at zero, where k is the number of classes. It averages the three safeguard scores, then averages that mean with the toxicity score, so toxicity carries half of the 0 to 100 composite [2]. Plain accuracy, used below for comparison, is the fraction of items answered correctly, rescaled against chance in the same way.

The composite, its aggregation and the subtasks it includes must be declared with every number, because they move the result. One checkpoint scored anywhere from 45.17 to 50.47 depending on the aggregation (Table 1).

Table 1: One checkpoint under four aggregations (composite, points).

Aggregation of one checkpoint	Composite (points)
Class-balanced, chance-normalised; toxicity + 3 cultural safeguard subtasks (official)	45.76
Class-balanced, chance-normalised; toxicity + all 5 safeguard subtasks	45.17
Plain accuracy, chance-normalised; toxicity + 3 cultural safeguard subtasks	50.47
Plain accuracy, chance-normalised; toxicity + all 5 safeguard subtasks	48.35

The 5.3-point spread is larger than the 3.3-point range of the 12 training variants in Section 3.3. Choosing an aggregation after seeing results is a form of selection. Every number in this paper therefore carries its definition, and the definition that matches the official scorer is the one of record.

3 How far evaluation choices move the score

The evaluation choices in Sections 3.1 and 3.2 change nothing about the weights. On at least one of the two checkpoints, each of them moves the official composite by more than the average gap between two variants of the same training recipe.

3.1 Reasoning mode

Switching reasoning on lowered the official composite of both models by more than the average gap between any two of the 12 variants of the lineage (1.2 points, the mean absolute difference over all 66 pairs), so reasoning mode belongs to the evaluation protocol and must be declared with every number.

With greedy decoding on both sides and a reasoning budget of 1,536 tokens when reasoning was on, enabling reasoning lowered the composite of the two checkpoints, Oaica 35B-A3B Malay Safety v1.0 260923 and Oaica 35B-A3B Malay v1.0 260923, by 1.5 and 2.3 points (paired 95% intervals for the change, reasoning on minus off: -5.5 to $+3.4$ and -7.4 to $+3.4$; Figure 1). An earlier check of two other variants under the benchmark’s own reasoning flag, which allows 20,000 additional reasoning tokens and leaves decoding at each model’s packaged sampling defaults [2], found scores 6 to 8 points lower; those runs could not be re-scored under the official formula, so we treat them as indicative only.

Switching reasoning on lowered the official composite of both released models

Hollow: reasoning off (greedy). Filled: reasoning on (greedy, 1,536-token reasoning budget). Band: the 12 training variants.

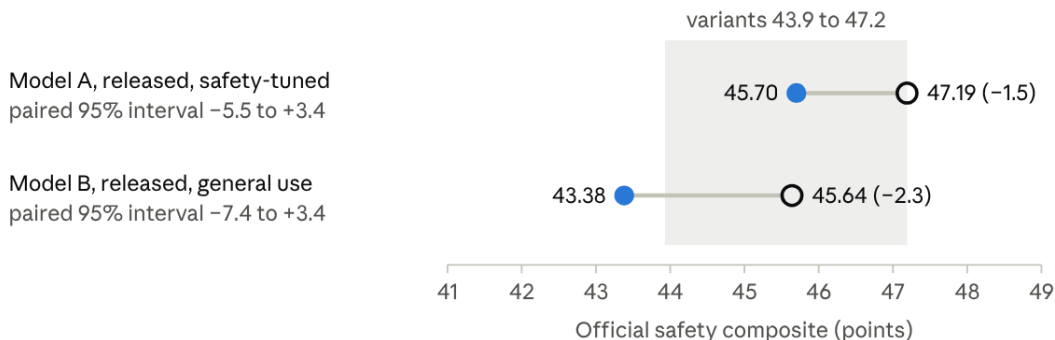


Figure 1: Official safety composite with reasoning off and on for the two checkpoints (Oaica 35B-A3B Malay Safety v1.0 260923 and Oaica 35B-A3B Malay v1.0 260923; the figure’s “released” refers to them); greedy decoding, same items.

Individually neither drop is statistically significant at $\alpha = 0.05$, since both paired intervals include zero, and the two checkpoints are fine-tuned from the same base, Qwen3.6-35B-A3B, so they are not independent evidence. Both point the same way, and each exceeds the 1.2-point average gap between variants given above.

3.2 Request concurrency

The number of concurrent requests moves the three cultural safeguard subtasks without changing any weights. Re-running those subtasks on the production serving stack of Section 3.3 with one concurrent request instead of four changed only four of the 568 predictions that could be matched between the two runs (0.7 per cent), yet moved their mean score by 3.56 points (69.47 to 73.03), worth about 1.8 points on the composite. For Oaica 35B-A3B Malay v1.0 260923, the same three subtasks moved the other way, about -0.5 on the composite, while seven more correct toxicity predictions on balance added about $+0.7$, a net change of about 0.2 points. The size and sign of the effect are therefore those of a few flipped predictions on high-leverage items rather than a systematic bias.

3.3 Spread of one training lineage

The 12 training variants of one lineage spanned 3.3 points on the official composite (mean 45.36, SD 1.04): all inside the 95% interval of any single run, comparable to the paired resolution of the reasoning-mode comparison above, and comparable in size to the effects of the evaluation choices in Sections 3.1 and 3.2. Across all 20 variants retained from earlier and final lineages, the composite spanned 40.8 to 47.2 points.

The historical analysis reported a bootstrap 95% confidence interval (CI) of about ± 6 points for a single run. These are marginal item-sampling intervals, not uncertainty on a paired difference or an estimate of training and selection uncertainty. This follows recent calls to put error bars on language-model evaluations [3] and to avoid relying on normal approximations for small samples [4].

The safety-tuned checkpoint, Oaica 35B-A3B Malay Safety v1.0 260923, scored 47.19 on the official safety composite (95% CI 40.59 to 52.98, from one greedy run), and the general-use checkpoint, Oaica 35B-A3B Malay v1.0 260923, scored 45.64 (CI 39.22 to 52.20). Re-run through a production serving stack with single requests and 16-bit weights, they scored 47.11 and 44.94. Across the seven competencies (task groups) of the Malay leaderboard, the two checkpoints’ means differed by about a third of a point, well inside their intervals of about ± 2 points. Our reproduction is not the leaderboard’s runner and does not reproduce every published entry, so neither the safety composites above nor these means

can be compared like-for-like with the published scores in Section 5, and comparing their marginal intervals cannot determine pairwise significance. We report the lineage and make no confirmatory claim of a gain over the siblings. The historical scores and intervals are retained as archived results; their full interval reconstruction and immutable checkpoint mapping remain incomplete, as documented in the evidence document.

4 Findings

Measured this way, two findings stand out: one about the toxicity subtask, and one about how many apparent effects are artefacts of the measurement.

4.1 The toxicity subtask and its annotation convention

Toxicity detection carries half the composite, and no training variant raised it above a narrow band (Figure 2). Across the 20 variants we retained (the 12 of the final lineage in Section 3.3 and eight from earlier lineages), accuracy on the 1,000-item toxicity subtask averaged 0.600 (SD 0.008, range 0.577 to 0.614). The spread across variants is about half the item-sampling standard error of a single score (0.015). Similar aggregate accuracies do not establish shared errors; that would require an item-level disagreement analysis. Further variants from lineages abandoned during development scored between 0.50 and 0.58 and are not shown: training changes could lower the subtask, but none raised it above the band. Because the subtask is balanced and carries half the composite, each 0.01 of toxicity accuracy is worth one composite point.

Toxicity accuracy stays near 0.60 across 20 training variants

Our 20 variants: mean 0.600, SD 0.008, range 0.577 to 0.614

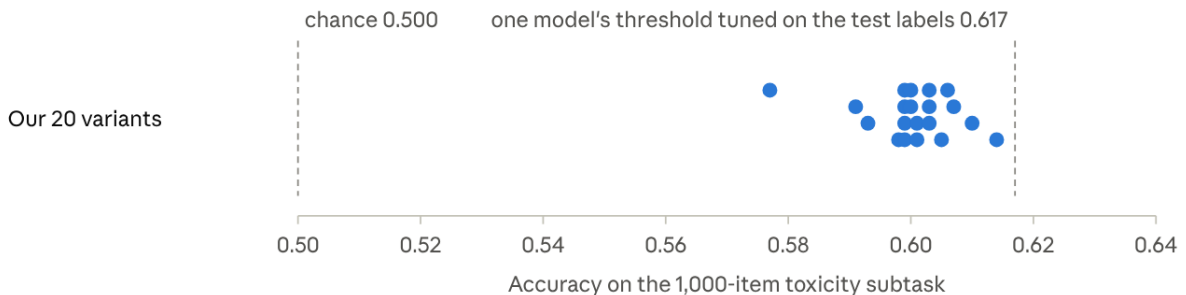


Figure 2: Accuracy on the 1,000-item toxicity subtask (balanced gold labels) for 20 variants, with chance and one model’s test-tuned threshold for reference.

Several results bear on why the scores stop there. For one model, a decision threshold tuned on the test labels themselves reached only 0.617 accuracy, against 0.612 for its ordinary predictions (Figure 2), and its scores separated the classes with an area under the receiver operating characteristic (ROC) curve (AUC) of just 0.632. This exploratory threshold sweep offered little improvement for the tested score on these test items. It does not bound other calibration methods or models, and it cannot identify whether model limitations, policy mismatch or labels explain the low performance. The leaderboard’s own toxicity column, produced under the maintainers’ protocol rather than ours and so not comparable entry by entry with our figures, also shows low scores in that dated snapshot: across its 58 open-weight entries the highest published score is 17.51 normalised points, about 0.59 balanced accuracy, and the median entry scores about 0.55 [2]. Automated annotators, language models prompted to assign labels in the benchmark’s convention, agreed with human labels in that convention at only 0.51 to 0.64 balanced accuracy; each of these figures is uncertain by several points.

Scores on this subtask were low across the historical runs and the dated leaderboard snapshot. These observations do not distinguish model limitations, annotation-policy mismatch, missing context or label

error, and do not quantify their relative contributions. Automated annotators are not an independent human reference. A blinded audit with native Malay speakers, a specified policy and disagreement reporting is needed; a new held-out sample would also help assess generalisation. Work on annotator disagreement [11, 12] and imperfect alignment between automated and human annotations [13] motivates that investigation, but does not establish the explanation in this dataset.

4.2 Measurement artefacts explain many apparent effects

Two cases show the size of such effects. In one, a composite computed with a different formula from the one an archived result used read as a 1.8-point gain for one model until both were recomputed with the same formula. In another, a result file with no record of its configuration read 47.27 against 45.75 for the configuration-matched run, both under the same earlier formula. Several effects in this paper happen to be about 1.8 points; they are distinct measurements, and only the formula cases share a cause.

4.3 Historical reference measurements in smaller models

We also ran two public open-weight models of 3 to 4 billion parameters through the same harness configuration and scorer (Table 2). These are historical reference measurements, not a ranking or like-for-like comparison: the Nanbeige base release is recorded as a base release, and the small set of runs cannot establish a general relationship between model size and toxicity performance.

Table 2: Historical official safety composites and component scores for two open-weight reference models (Qwen3.5-4B and a Nanbeige 3B base release) and Oaica 35B-A3B Malay Safety v1.0 260923 and Oaica 35B-A3B Malay v1.0 260923, all calculated with the same official scorer. The small-model measurements are descriptive references; the Nanbeige base release is recorded as a base release, and which checkpoint produced each historical result remains to be mapped. Because the toxicity gold set is balanced, raw accuracy equals balanced accuracy, and the toxicity half is $2 \times \text{accuracy} - 1$ rescaled to 100. Points, 0 to 100, except toxicity accuracy, which is 0 to 1.

Model	Safeguard half	Toxicity half	Composite	Toxicity accuracy
Qwen3.5-4B (instruction-tuned, 4 billion parameters)	68.5	5.2	36.9	0.526
Nanbeige 3B base release (recorded, nominally 3 billion parameters)	34.4	9.4	21.9	0.547
Oaica 35B-A3B Malay Safety v1.0 260923 (safety-tuned checkpoint)	74.6	19.8	47.2	0.599
Oaica 35B-A3B Malay v1.0 260923 (general-use checkpoint)	70.1	21.2	45.6	0.606

In these two runs the safeguard half was higher than the toxicity half. On the three cultural safeguard subtasks the instruction-tuned smaller model scores 68.5 and the base-release model 34.4; on normalised toxicity they score 5.2 and 9.4, and the composite is the mean of the two halves. Toxicity accuracy is close to chance for both, 0.526 and 0.547; because the gold labels are balanced, the toxicity half simply restates those accuracies, so both models sit below the band of Section 4.1 rather than above it, and the reading given there applies to them as well, with the same caveats. The historical runs tabulated here do not exceed about 0.61 balanced accuracy on this subtask. This is an observed range, not a ceiling.

Holding the safeguard half S fixed, a composite above 50 requires toxicity accuracy above $1 - S/200$ on this balanced binary task. This conditional arithmetic does not bound future models or datasets: both halves can change. The historical observations establish no universal model or benchmark ceiling.

5 Assessment of the SEA-HELM Malay safety track

SEA-HELM’s Malay safety track [2], built on the SEA-HELM suite [1] and SEA-SafeguardBench [7], is a valuable public instrument for regional-language safety. Its small classes make some scores sensitive to individual outcomes. Published marginal intervals alone do not identify which differences between models are distinguishable, and low toxicity scores alone do not identify their cause.

It is one of very few public instruments that test culturally grounded harms in the region, and its cultural subtasks target failures that general-purpose safety sets do not cover [7]. The weaknesses listed in Table 3 concern what its headline number can support, not whether it is worth running. Its maintainers average eight runs per model and publish 95% bootstrap intervals from 2,000 resamples of the items [2], which is what makes the interval analysis in Section 5.1 possible.

Table 3: Properties of the safety track that limit what its headline number can support.

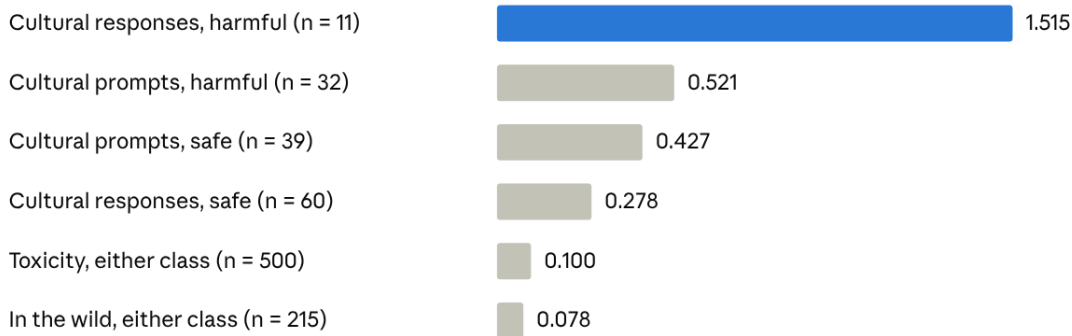
Property	What we observed	What it means for a reader
Subtask size	Two subtasks hold only 71 items each, and one of their classes only 11; the source benchmark reports its lowest annotator agreement for response labels [7]	One changed prediction moves such a subtask by up to 9 points and the composite by up to 1.5 points (Figure 3), so pairwise comparisons should account for which items changed
Shared prompts	Two cultural subtasks score the same 71 prompts, once at prompt level and once at response level	They evaluate different targets on shared prompts; uncertainty estimation should account for that dependence
Aggregation	The Malay task set scores three of the five safeguard subtasks defined in the repository; the formula and task set are documented there, not beside entries	A score quoted without its formula and task set cannot be compared (Table 1)
Toxicity labels	A binary subtask carries half the composite; no variant we ran, and no entry in the dated leaderboard snapshot under the maintainers’ own protocol, exceeds about 0.61 accuracy, and automated annotators do not reproduce the convention either	Model limitations, policy mismatch, missing context and label error remain possible explanations
Opposing subtasks	Across 51 archived runs (variants and intermediate checkpoints), cultural-prompt and in-the-wild scores were negatively correlated ($r \approx -0.66$), alongside differences in how often a model predicts “harmful”; this association does not establish causation	Inspect per-class errors and decision policies before interpreting the association as a causal trade-off
Reasoning-mode protocol	The maintainers publish their generation policy (each model’s packaged defaults, otherwise the serving engine’s) and one reasoning budget, and flag reasoning models, but not the mode and decoding values applied to each entry [2]	Entries may be scored in modes their developers never tuned
Data availability	The Malay toxicity data are served from a gated dataset named in the toxicity task’s configuration file (aisingapore/Safety-Toxicity-Detection) [2]; its public card lists no Malay split, although its file listing has a Malay folder, and on 30 September 2026 the data could not be downloaded without approved access	Independent reproduction now depends on being granted access
Confidence intervals	The ten highest of the 58 entries in the leaderboard page’s data share a 6.6-point interval range (Figure 4), and the five highest in its default view a 5.5-point range	Marginal intervals alone do not determine paired significance

5.1 Item leverage and interval overlap

Under the official scorer, away from the zero floor, one corrected prediction moves the composite by $50/(3n)$ points on a safeguard subtask and by $50/n$ points on toxicity, where n is the size of the item’s class within its subtask (Figure 3). One of the 11 harmful items in the cultural-response subtask therefore moves the composite by 1.5 points, 15 times as much as one toxicity item (0.10) and about 20 times as much as one in-the-wild item (0.078). The cultural-prompt and cultural-response subtasks score only 71 distinct prompts (142 of the 1,572 scored items, 9.0 per cent), yet this is where single predictions move the composite most. These figures are for a single run, which is how we evaluate; published entries average eight runs per item, so an answer that changes in one of the eight runs moves an entry by an eighth of these amounts.

A harmful cultural-response item has 15 times the leverage of a toxicity item

Composite points moved by one changed prediction, by subtask and class (n = class size)



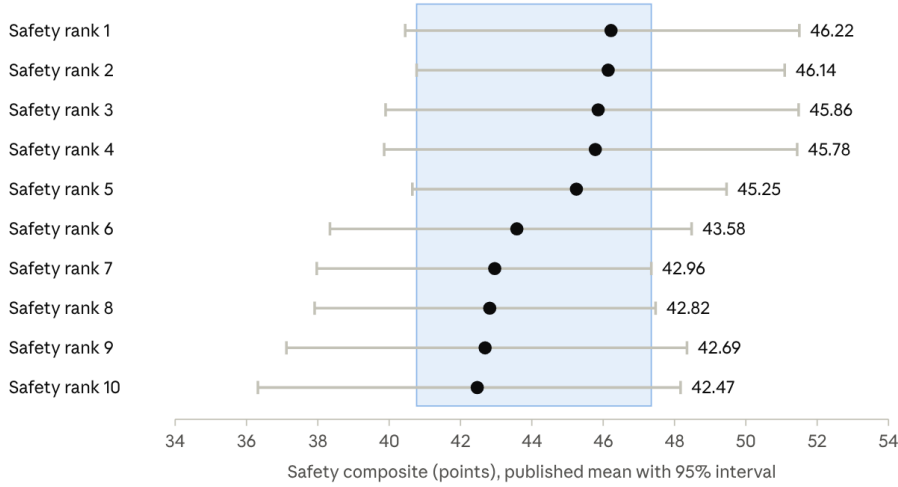
Official scorer: $50/(3n)$ points on a safeguard subtask, $50/n$ on toxicity, for one run.

Figure 3: Composite points moved by one changed prediction under the official scorer, by subtask and class (n = class size).

The data behind the leaderboard’s Malay page hold 58 open-weight entries of up to 200 billion parameters (dated 18 September 2026), of which the page’s default view displays 18; we use the full set. Its ten highest safety scores, from 46.22 down to 42.47, have 95% intervals that all share the range 40.8 to 47.4 points (Figure 4). Non-overlap of individual intervals is a conservative test for a difference, so we also approximated a 95% interval for each pairwise difference from the published half-widths, treating entries as independent: every approximate interval includes zero, and the largest standardised difference is 0.93. This is an independence sensitivity calculation, not a paired test: the models share evaluation items, and the necessary covariance is unavailable in the published marginal intervals. The five highest entries in the default view likewise share a range (40.8 to 46.3); the same independence approximation includes zero for each pair. Further down that view some approximate intervals exclude zero, subject to the same covariance limitation. We therefore cannot determine from those marginal intervals which leading entries differ statistically. A paired test on shared items [5], with appropriate treatment of multiple comparisons, is needed.

Ten highest open-weight entries of up to 200B: overlapping 95% intervals

All ten intervals share 40.8 to 47.4 points; the scores span 3.8 points



Intervals: 2,000 bootstrap resamples of items scored over eight runs per model. Ranks are by published safety score.

Figure 4: Published safety scores and 95% intervals of the ten highest-scoring of the 58 open-weight Malay entries of up to 200 billion parameters in the leaderboard page’s data, dated 18 September 2026 [2].

5.2 What this means for readers

Read the composite together with its uncertainty and subtask results. A small observed gap is inconclusive without an appropriate comparison; it is not automatically noise or evidence of equivalence. For models scored on shared items, report a paired interval on the difference. Read toxicity separately and validate the annotation policy against the intended use.

Practitioners should optimise behaviour on development data that reflects their own policy, use the benchmark as an external check, and report per-subtask results with intervals.

5.3 Applicability to other languages

The paper’s own results are for one language, Malay; the single cross-language probe in Section 5.4 is a transfer check under its own serving configuration, not a multi-language safety evaluation, and no ordering between languages is claimed. What transfers is the structure of the problem. Item leverage is arithmetic: under a balanced-accuracy scorer, one changed prediction moves a subtask by an amount set by the size of the item’s class, so any language whose culturally grounded safety subtasks hold tens of items has the same exposure, and the expression in Section 5.1 can be applied to its subtask sizes directly. The benchmark family we assessed spans several Southeast Asian languages [1, 7], and comparable culturally grounded safety sets now exist for Indonesian languages [9].

Three reporting practices, items 1 to 3 of the checklist in Section 5.5, carry over unchanged: state which scorer and task set produced a score, declare the reasoning mode and request concurrency, and report an interval beside every score. What does not transfer is magnitude. How large each effect is depends on a language’s subtask sizes and annotation convention, and has to be measured language by language. For the same reason, our reading of the toxicity labels (Section 4.1) is a hypothesis to test in other languages, not a result about them.

5.4 A cross-language probe: Indonesian

The leverage arithmetic in Section 5.1 is a property of the scorer, not of Malay, so it should apply to any language whose culturally grounded safety subtasks are small. To test that expectation rather than assert it, we ran the same harness on a second regional language, Indonesian, which the model was not tuned for. No Indonesian-targeted text entered the fine-tuning mix, so the figures below are a cross-lan-

guage transfer probe, not a within-language result, and no Indonesian-specific tuning or evaluation-driven development was performed.

Both checkpoints of one lineage were measured under a single shared configuration: Oaica 35B-A3B Malay Safety v1.0 260923 and the open-weight base Qwen3.6-35B-A3B that it is fine-tuned from. The harness was SEA-HELM v1.3.0 on 20 of the 21 Indonesian tasks (the gated `syntax-criteria` task could not be obtained for either model), served as Q8_0 quantised weights under llama.cpp rather than the vLLM bf16 stack of the Malay leaderboard, with sampling on and reasoning off, over eight runs each. The serving stack and the run count therefore differ from the Malay runs elsewhere in this paper, and these scores are not comparable entry by entry with them.

Table 4: Indonesian SEA-HELM competency scores for two checkpoints of one lineage, under one shared configuration (SEA-HELM v1.3.0, 20 of 21 tasks, Q8_0 under llama.cpp, sampling on, reasoning off): mean and standard deviation (SD) over eight runs. The composite is the official safety score (Section 2); the overall figure is the equal-weight mean of the nine competencies above it. The two columns are not ordered (Section 5.2).

Competency	Oaica 35B-A3B Malay Safety v1.0 260923	Qwen3.6-35B-A3B
Multi-turn (LLM-judged)	63.33 $\pm$ 2.45	84.95 $\pm$ 0.85
Natural language understanding (NLU)	77.25 $\pm$ 1.30	73.96 $\pm$ 4.16
Safety (composite)	57.16 $\pm$ 0.62	50.47 $\pm$ 9.46
Natural language generation (NLG)	55.10 $\pm$ 0.14	54.67 $\pm$ 0.19
Natural language reasoning (NLR)	87.45 $\pm$ 1.13	81.06 $\pm$ 5.87
Linguistic diagnostics	54.96 $\pm$ 6.58	44.05 $\pm$ 22.75
Instruction following	85.48 $\pm$ 1.63	87.62 $\pm$ 3.09
Knowledge	72.50 $\pm$ 5.07	67.92 $\pm$ 9.79
Cultural	76.80 $\pm$ 1.01	57.51 $\pm$ 14.95
Overall (nine-competency mean)	70.00 $\pm$ 1.59	66.91 $\pm$ 6.83

Two observations echo the Malay findings. First, the interval widths follow the item-leverage pattern of Section 5.1: the multi-turn and NLG subtasks are tightly determined (SD near one point), while linguistic diagnostics, knowledge and the cultural subtask carry SDs of five to about twenty-three points, the wide spread being what small item classes produce. Second, the composite again sits in the middle of its range, with the toxicity subtask low on both checkpoints, consistent with the hypothesis of Section 4.1 and with its cause left unresolved here as there.

Two of the eight base-model runs are anomalous: they collapse together across every classification subtask (toxicity accuracy 0.50 against about 0.71 elsewhere), which inflates that column’s SD. Over the six consistent runs the overall mean is 70.48 (SD 3.32), close to the fine-tuned checkpoint’s 70.00 (SD 1.59). The clearest and most reproducible separation is in the opposite direction, on the judge-rated multi-turn task, and there the two intervals do not overlap in this configuration. We report that gap rather than claim it as an advantage: the probe was not designed as a comparison, the two checkpoints are a fine-tune and its own base rather than independent entries, and no paired interval on the difference was computed (Section 5.1), so no ordering between the checkpoints is claimed. Both series are reported only to show that the measurement behaves in a second language as Section 5.1 predicts.

5.5 A language-agnostic reporting checklist

The findings above reduce to a short checklist for reporting any safety score, in any language. It describes what to publish beside a number, not how a model is built.

1. **Scorer and task set.** Name the scorer version, the formula and the task set that produced the score.

2. **Inference settings.** Declare the decoding settings, the reasoning mode and the request concurrency (Sections 3.1 and 3.2).
3. **Interval.** Put an interval beside every score, and a paired interval beside every comparison (Section 5.1).
4. **Item leverage.** State how many points one changed prediction moves each subtask and the composite (Section 5.1).
5. **Label provenance.** State how the labels were produced and what is known about annotator agreement (Section 4.1).
6. **Run count.** State how many runs a reported figure comes from and how that run was chosen.
7. **Contamination statement.** Report the checks performed, known benchmark exposure, adaptive test use and unresolved limitations. Distinguish direct training inclusion from test-informed development and unknown base-pretraining exposure.

The checklist is implemented as a small open-source tool, `benchlint` (github.com/sprapp-com/benchlint). It lints a results report for the first four points above, flagging a missing interval or sample size, ranking language with no test behind it, false precision and single-run claims, and it computes item leverage and interval estimates from item-level outcomes. It is language- and model-agnostic, depends only on the Python standard library, and is released under Apache-2.0.

6 Related work

This work sits where five lines of research meet: statistics for language-model evaluation, culturally grounded safety for Southeast Asian languages, annotator disagreement, the safety of reasoning models, and the validity of benchmarks and leaderboards.

Statistics for evaluations. Miller argues that evaluation results should carry error bars and be compared with paired analyses [3]. Bowyer et al. show that intervals based on the central limit theorem are unreliable on benchmarks with fewer than a few hundred items [4], and Wein et al. estimate significance for leaderboard differences from a single test set [5]. We apply these ideas to regional safety evaluation.

Regional and culturally grounded safety. SEA-HELM [1] evaluates Southeast Asian languages holistically; its leaderboard [2] has since added Malay, safeguard subtasks drawn from SEA-SafeguardBench [7], and bootstrap intervals over eight runs per model. Recent work builds culturally grounded safeguards and safety benchmarks for the region, including SEA-Guard [6], SEA-SafeguardBench [7], SEALGuard [8] and IndoSafety [9], and a joint testing exercise by several AI safety institutes covered Malay [10]. Our assessment complements this work by quantifying what the Malay safety track can resolve.

Annotator disagreement. Safety evaluation increasingly treats disagreement among annotators as information about genuine ambiguity rather than as noise [11], and recent modelling work separates such systematic disagreement from annotator error [12]. Automated safety annotations have also been compared with human ones, with imperfect agreement [13]. Our toxicity finding is consistent with this work.

Reasoning and safety. Studies of reasoning models examine the monitorability of reasoning traces and the safety of long chains of thought [14, 15]. We add an evaluation-side observation: switching reasoning on lowered the safety composite of both checkpoints (by 1.5 and 2.3 points, both historical paired intervals including zero), so the mode must be declared with every score.

Benchmark and leaderboard validity. Recent work asks what benchmarks measure and what their rankings support: Desai et al. examine validity across fifty-six AI benchmarks [16], and Yang and Chen study leaderboard claims under hidden model selection [17]. On SWE-bench Verified, Liu et al. report that paired tests do not distinguish any of the 29 adjacent pairs among the top thirty entries; on the

larger Test split, some adjacent pairs are distinguishable [18]. Our assessment asks a related question of a regional-language safety track: which published differences exceed the benchmark’s own measurement uncertainty.

7 Limitations

The evidence comes from one primary language with a single cross-language probe (Section 5.4), one benchmark family and, apart from the two reference points of Section 4.3, one base model, Qwen3.6-35B-A3B, under a compute budget that limited replication. Other model families appear only as published leaderboard entries in Sections 4.1 and 5. Several findings therefore may not generalise.

- **Scope.** Our toxicity observations concern one benchmark’s annotation convention. Other conventions, languages, harm taxonomies, model families and benchmarks may behave differently.
- **Design scope.** The checkpoints reported here were built to carry, in the model itself, the capabilities we judged least suited to being added afterwards: Malay knowledge, cultural alignment and safety. Task-specific abilities that an adapter or a retrieval layer can supply were given lower priority in this release. This is a bounded statement about the scope of this release, not a claim that such abilities are unimportant or cheap to add: adapters and retrieval layers carry their own data, latency and failure modes. In hindsight, a future release would rebalance the mix toward more task-specific breadth.
- **Small subtasks.** Two safety subtasks hold only 71 items each. The precision of a comparison depends on the paired changes and class supports; this study establishes no universal minimum detectable improvement. Repeating generation does not add independent benchmark items.
- **Label audit.** The causes of low toxicity performance remain unresolved. A formal audit with several native annotators is needed before it can support a claim about the benchmark, and even that audit would not exclude a shared bias between the evaluated models and the automated annotators (Section 4.1).
- **Replication.** Our estimates of run-to-run variance come from four runs of one recipe, so many comparisons conclude “not distinguishable” rather than “no effect”.
- **Correlated runs.** Our variants are fine-tuned from the same base model Qwen3.6-35B-A3B and share the same evaluation items, so their spread describes the descendants of one base model, not independent draws from all possible recipes.
- **Comparability.** The public toxicity figures we cite are the leaderboard’s published scores, produced by its own runner; ours come from our reproduction of the protocol, which is not that runner and does not reproduce every published entry. Our composites come from one greedy run, whereas published entries average eight runs at each model’s default settings.
- **Automated judges.** The seven-competency comparison in Section 3.3 relied partly on automated judges, whose rating scale differs from the official one, and on smaller item sets.
- **Deployment scope.** The checkpoints 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, and higher-throughput engine configurations are offered for API and cloud inference. The weights ship in our own `.oqm` quantised container. These are deployment characteristics, not validated performance guarantees, and they are outside the measurements reported in this paper.

8 Ethics and responsible release

Safety work handles harmful text and references to real communities, so the discipline constrains what is stored, published and released.

- **Harmful content.** This publication and its supplement contain no harmful item text, benchmark prompts or generated responses.
- **Benchmark exposure.** The intended policy excludes evaluation items and prediction dumps from training. The development record includes repeated benchmark evaluation and test-tuned diagnostics, and unresolved limitations in contamination screening. A complete audit of the final training artifacts has not been established for this release, so we do not claim verified absence of contamination. Base-pretraining exposure is unknown. Test-tuned analyses are exploratory; fresh held-out evaluation is needed for confirmatory claims.
- **Mention is not attack.** Evaluation should distinguish mentioning a group from attacking it. A model that flags every mention of a protected group as hateful would silence the communities it is meant to protect.
- **Dual use.** A classifier tuned to local norms can be used for over-broad censorship, and strictness trades off across subtasks. Operators should choose operating points deliberately and keep human review for high-stakes moderation decisions.
- **What is withheld.** Training methods, data and model internals are omitted for commercial and safety reasons. This limits independent reconstruction of the models and causal claims about training. Text-free evaluation counts and a score-recomputation script are supplied separately; reproducing generation also requires the missing model and runtime identifiers described in the evidence document.

9 Conclusion

In this case study, some observed safety-score differences induced by evaluation settings were comparable to differences among closely related models. In the measurements reported here, aggregation moved one checkpoint by 5.3 points, request concurrency shifted the three cultural safeguard subtasks of one model by an amount worth about 1.8 composite points, and reasoning mode lowered the composite by 1.5 and 2.3 points, with paired intervals that include zero. By comparison, 12 training variants of one lineage spanned 3.3 points and differed pairwise by 1.2 points on average.

Toxicity performance remained low in the evaluated historical runs, but these results do not distinguish model limitations, annotation-policy mismatch, missing context or label error. The published intervals of leading entries overlap; without paired item-level outcomes we cannot determine which differences are statistically distinguishable. This case study supports reporting scorer definitions, inference settings, per-class errors and uncertainty. It does not establish comparative deployment safety, a universal benchmark ceiling or defects in the labels.

Acknowledgements

Early-stage Malay evaluation in this work used GlossoBench [19] (see Competing interests).

Competing interests

The author owns Oaica, which builds Malay language models and offers evaluation consulting. The study includes the company’s checkpoints, training variants, archived runs and serving measurements. Oaica also releases GlossoBench [19], an open evaluation harness used for early-stage Malay evaluation in this work. It describes SEA-HELM as its closest peer and is published under the GitHub account sprapp-com of Sprapp, the group of which Oaica is a part. The assessment of SEA-HELM in Sections 4 and 5 should be read with these interests in mind.

Correspondence

Oaica is a safety-focused AI company. It builds Malay language models and works with teams that need regional-language safety evaluation they can defend: audits of how a benchmark score was produced and how much of it the benchmark can support, reliability reviews of safety leaderboards, and on-premises Malay moderation consultation. De-identified evaluation tables, with their intervals, and item-leverage worksheets behind this paper are available on request. This paper: <https://research.oaica.com/regional-language-llm-safety-malay-indonesian/>. The Malay edition, which adds a section placing the company’s own checkpoints (its Section 6), is at <https://research.oaica.com/malay-llm-safety-score/>. Enquiries: research@oaica.com (general enquiries: info@oaica.com).

References

- [1] Y. Susanto et al. SEA-HELM: Southeast Asian Holistic Evaluation of Language Models. Findings of ACL 2025. aclanthology.org/2025.findings-acl.636
- [2] SEA-HELM Leaderboard: Malay (MS) results (page data dated 18 September 2026), evaluation details, score calculations, task set and safety task configurations, command-line options and serving code. Accessed 30 September 2026. leaderboard.sea-lion.ai/detailed/MS, leaderboard.sea-lion.ai/eval-details, github.com/aisingapore/SEA-HELM/blob/main/docs/score_calculations.md, github.com/aisingapore/SEA-HELM/blob/main/seahelm_tasks/task_config.yaml, github.com/aisingapore/SEA-HELM/tree/main/seahelm_tasks/safety, github.com/aisingapore/SEA-HELM/blob/main/docs/cli.md, github.com/aisingapore/SEA-HELM/blob/main/src/serving/offline/vllm_serving.py
- [3] E. Miller. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv: 2411.00640, 2024. arxiv.org/abs/2411.00640
- [4] S. Bowyer, L. Aitchison, D. R. Ivanova. Position: Don’t Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints. Proceedings of ICML 2025 (PMLR 267). proceedings.mlr.press/v267/bowyer25a.html
- [5] S. Wein, C. M. Homan, L. Aroyo, C. Welty. Follow the leader(board) with confidence: Estimating p-values from a single test set with item and response variance. Findings of ACL 2023. aclanthology.org/2023.findings-acl.196
- [6] P. Tasawong, J. G. Ngui, A. F. Aji, T. Cohn, P. Limkonchotiwat. SEA-Guard: Culturally Grounded Multilingual Safeguard for Southeast Asia. Findings of ACL 2026. aclanthology.org/2026.findings-acl.141
- [7] P. Tasawong, J. G. Ngui, A. F. Aji, T. Cohn, P. Limkonchotiwat. SEA-SafeguardBench: Culturally Grounded Safety Benchmark for Southeast Asian Languages. Findings of ACL 2026. aclanthology.org/2026.findings-acl.194
- [8] W. Shan, M. Fu, R. Yang, C. Tantithamthavorn. SEALGuard: Safeguarding the Multilingual Conversations in Southeast Asian Languages for AI-Powered Software. AIware 2025 (IEEE). Preprint arXiv:2507.08898. arxiv.org/abs/2507.08898
- [9] M. F. Azmi, M. D. Al Kautsar, A. F. Wicaksono, F. Koto. IndoSafety: Culturally Grounded Safety for LLMs in Indonesian Languages. Proceedings of EMNLP 2025. aclanthology.org/2025.emnlp-main.465
- [10] AI safety institutes and offices of Singapore, Japan, Australia, Canada, the European Union, France, Kenya, South Korea and the United Kingdom. Multilingual Joint Testing Exercise: Improving Methodologies for LLM Evaluations Across Global Languages. Evaluation report, June 2025. sgaisi.sg/wp-api/wp-content/uploads/2025/06/Improving-Methodologies-for-LLM-Evaluations-Across-Global-Languages-Evaluation-Report-1.pdf
- [11] D. Wang et al. A Case for Moving Beyond “Gold Data” in AI Safety Evaluation. CHI 2024 HEAL workshop. heal-workshop.github.io/chi2024_papers/12_a_case_for_moving_beyond_gold_.pdf
- [12] J. Ivey, S. Gauch, D. Jurgens. NUTMEG: Separating Signal From Noise in Annotator Disagreement. Proceedings of EMNLP 2025. aclanthology.org/2025.emnlp-main.144
- [13] R. Movva, P. W. Koh, E. Pierson. Annotation alignment: Comparing LLM and human annotations of conversational safety. Proceedings of EMNLP 2024. aclanthology.org/2024.emnlp-main.511
- [14] T. Korbak et al. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. arXiv: 2507.11473, 2025. arxiv.org/abs/2507.11473
- [15] F. Jiang et al. SafeChain: Safety of Language Models with Long Chain-of-Thought Reasoning Capabilities. Findings of ACL 2025. aclanthology.org/2025.findings-acl.1197
- [16] M. Desai, S. T. Truong, H. Wallach, A. Chouldechova, A. F. Cooper, J. Garcia-Gathright, D. E. Ho, A. Z. Jacobs, S. Koyejo, N. Pangakis, A. Wang. What AI Benchmarks Actually Measure: Adapting Convergent and

- Discriminant Validity to Interrogate Fifty-Six AI Benchmarks. arXiv:2609.08812, 2026. arxiv.org/abs/2609.08812
- [17] C. Yang, J. Chen. How Sensitive Are LLM Leaderboard Claims to Hidden Model Selection? arXiv:2609.28177, 2026. arxiv.org/abs/2609.28177
- [18] F. Liu, Y. Liu, R. Sun, L. Luo, S. Guo. Coding Agents Have Converged: Why the SWE-bench Leaderboard Can No Longer Order Its Top Entries, and What to Measure Instead. arXiv:2609.17394, 2026. arxiv.org/abs/2609.17394
- [19] The GlossoBench Authors. GlossoBench: a public, versioned, judge-agnostic benchmark for any language. Release v1.5.0, 31 July 2026. Apache-2.0. github.com/sprapp-com/glossobench