OUTBACK YAK RESEARCH · WHITEPAPER · V0.2

NepNewsCluster

A bilingual Nepali–English news consolidation benchmark across thirteen frontier and open-weight large language models.

May 2026 · 107 stories · 1,310 articles · 23 publishers · 2 generation runs

13 MODELS EVALUATED	107 CLUSTER QUESTIONS
1,310 ARTICLE SNIPPETS	23 NEPALI PUBLISHERS

§ 01

EXECUTIVE SUMMARY

Nepali (~32 million speakers) is well-served by classification benchmarks but absent from bilingual generation evaluation. NepNewsCluster fills that gap with a single, narrowly-scoped, deeply-instrumented test: synthesise multiple news articles into one bilingual brief, measured by an LLM judge against a rubric reverse-engineered from human grading behaviour.

This is the first multi-document, bilingual, rubric-graded news consolidation benchmark for Nepali that we are aware of. Each of 107 cluster-questions packages up to fifteen real articles from different Nepali outlets — covering one news story — and asks the model to produce a Nepali-Devanagari headline, an English headline, a 3–4 sentence summary in each language, and a typed list of named entities mentioned in the English summary. Outputs were graded blind on three axes (Nepali prose, English prose, topic coverage), each scored 1–10 by Claude Opus 4.7 with extended thinking. We report results scaled to a familiar /100 axis-quality scale throughout.

HEADLINE RESULT

81/100

Claude Sonnet 4.6 wins both runs at 81 mean axis quality, ahead of every other model on every axis.

REASONING SURPRISE

+21 / −3

Test-time reasoning gave **opposite signs** for the two models tested directly — Qwen 3.6 Max think gained 21 points, DeepSeek V4 Pro think lost 3.

LOCAL-RUNNABLE COMPETITIVE

$0.0005

Gemma 4 31B beats GPT-5.4 mini and Claude Haiku 4.5 at less than 1/40th the per-call cost.

TOP RANKS ARE DURABLE

2 / 2

Across two independent generation runs and a tightened rubric, **top-3 ranks are stable**. Mid-table reshuffles meaningfully.

BUCKET-LEVEL EFFECTS

WHERE REASONING BREAKS

2.7 pts

Local <150B trails Chinese SOTA by 4.3 axis points; Chinese SOTA trails US SOTA by just 1.1. The within-bucket spread is wider than the between-bucket gap.

−13

DeepSeek's worst regression: a thinking trace that synthesised arrest counts *across days no source ever combined*, then committed to the fabricated total.

§ 02

WHY THIS BENCHMARK

02

The Nepali NLP gap

Nepali is the lingua franca of Nepal and a substantial diaspora language across Australia, the Gulf, and South Asia. Despite an active research community at IRIIS Nepal, Kathmandu University ILPRL, Tribhuvan University, and NAAMII, the existing benchmark suites for Nepali are concentrated on classification and reading-comprehension tasks. None of them measure the things that matter most for a production news pipeline like K cha khabar: bilingual generation, multi-document synthesis, factuality across two scripts simultaneously.

EXISTING BENCHMARK	YEAR	TASK SHAPE	WHY THIS WORK DOESN'T OVERLAP
NLUE	2024	12 GLUE-style tasks (NER, sentiment, NLI, ...)	Discriminative only. No bilingual generation. No LLM-judge.
IndicGenBench	2024	Pan-Indic generation, ~500 Nepali items in CrossSum-In split	Single-article. ROUGE/chrF.
Belebele	2023	MCQ reading comprehension, Devanagari + Latin	Multiple choice, not generation.
Global-MMLU	2024	Translated MMLU across 42 languages incl. Nepali	Knowledge MCQ, not consolidation.
FLORES-200	2022	Translation pairs across 200 languages	Single-sentence translation, not generation under multi-source constraint.

The gap NepNewsCluster targets is a triangulation: **(a)** multi-document consolidation, **(b)** bilingual parallel generation with parallel grading, and **(c)** rubric-based LLM-judge methodology against realistic Nepali source distribution — production RSS feeds where outlets disagree on numbers and dates, sensational framings drift across the news cycle, and Bikram Sambat ↔ Gregorian conversions catch out even competent models.

> *A flawed first cut beats no cut. We publish this because, to our knowledge, no equivalent Nepali benchmark exists.*

Why K cha khabar built it

K cha khabar (kchakhabar.com) is an Outback Yak product: a cross-publisher Nepali news intelligence platform that ingests RSS feeds from thirty-plus Nepali outlets, classifies and clusters articles by underlying story, then renders each cluster as a bilingual brief with ownership transparency and cross-publisher echo detection. Bilingual consolidation is the central LLM-bound step in that pipeline. The economics of running it at production scale every fifteen minutes — for a low-resource language, on multilingual models that are not Nepali-tuned — depend entirely on which model performs the task well at low cost. Without a benchmark, model choice is anecdote.

NepNewsCluster is the answer to the question we already had to answer privately. Publishing it with this much methodological transparency is our contribution back to the Nepali NLP ecosystem.

§ 03

THE TASK

03

What models had to do

Given between three and fifteen short news articles from different Nepali outlets covering the same story, the model had to produce in a single completion:

- An **English headline** of twelve words or fewer
- A **Nepali (Devanagari) headline** of fifteen words or fewer — actual Nepali, not transliteration
- A **3–4 sentence English summary** covering the who/what/where/when/why
- A **3–4 sentence Nepali (Devanagari) summary** conveying the same facts in idiomatic Devanagari prose
- A **list of named entities** mentioned in the English summary, typed as `person` · `org` · `place` · `party` · `event` · `policy`
- A **URL slug** derived from the English headline

This is the same contract the K cha khabar production summariser fulfils on every cluster. The task is hard for three reasons that compound:

MULTI-DOCUMENT DISAGREEMENT

Outlets contradict each other on arrest counts, casualty numbers, dates, named titles, and political-party affiliation. The model has to commit to a shared truth without inventing one.

BILINGUAL INDEPENDENCE

Nepali and English can fail independently. A correct Nepali summary paired with a mistranslated English version (or vice-versa) is a failure mode that's invisible if you only score one language.

CALENDAR CONVERSION

Nepali outlets publish dates in Bikram Sambat. The model often has to convert BS↔Gregorian to align with English-language sourcing. Off-by-one-month errors are common.

The grading rubric reflects all three pressures, scoring each one explicitly and penalising both axes when an output internally contradicts itself.

§ 04

METHOD

Data, models, rubric, judge

04

Data construction

107 cluster-questions were sampled from a snapshot of approximately 9,000 active clusters in the K cha khabar production database, ranked by `COUNT(DISTINCT publisher_id)` — a stronger popularity proxy than raw article count. The sample was deliberately stratified rather than pure top-N to avoid overweighting any single coverage tier:

DISTINCT PUBLISHERS	CLUSTERS	RATIONALE
≥ 9 (high coverage)	47	"Major" stories — every outlet covered
8 publishers	10	Trimmed from 41 to avoid long-tail dominance
7 publishers	42	Mid-coverage
5 publishers	5	Mid-coverage sample
3 publishers	3	Low-coverage edge case
Total	**107**	

Each cluster contributed up to its fifteen most recent articles to the question. In aggregate the corpus spans **1,310 article snippets** (mean 12.2 per question, drawn from a pool of 1,782 in the unfiltered source clusters), **89% Nepali** / 11% English by article count, across **23 unique publishers**. The publisher list is reproduced in §10. Each article is presented to the model as `publisher | language | publishedAt | EN headline | NE headline | excerpt(≤280 chars) | tags` — exactly the shape the production summariser already sees. The K cha khabar production summary itself is *not* in the prompt, so the eval is genuinely a from-scratch consolidation for every model.

Model lineup (test-3, primary)

Thirteen models in three buckets: US frontier proprietary, Chinese frontier (open-weight, large), and locally-runnable (under 150B parameters, open-weight). Two models —

DeepSeek V4 Pro and Qwen 3.6 Max — appear as both think and no-think variants, controlling for nothing but the test-time reasoning toggle and the output-token budget.

MODEL	BUCKET	REASONING	INPUT $/M	OUTPUT $/M
GPT-5.4	US SOTA	off	2.50	15.00
GPT-5.4 mini	US SOTA	off	0.75	4.50
Claude Sonnet 4.6	US SOTA	off	3.00	15.00
Claude Haiku 4.5	US SOTA	off	1.00	5.00
DeepSeek V4 Pro (no-think) †	Chinese	off	0.435	0.87
DeepSeek V4 Pro (think) †	Chinese	on	0.435	0.87
GLM 5.1 (z.ai)	Chinese	off	1.05	3.50
Xiaomi MiMo V2.5 Pro	Chinese	off	1.00	3.00
Qwen 3.6 Plus	Chinese	off	0.325	1.95
Qwen 3.6 Max (no-think)	Chinese	off	1.30	7.80
Qwen 3.6 Max (think)	Chinese	on	1.30	7.80
Gemma 4 31B (it)	Local	off	0.13	0.38
Qwen 3.6 27B (no-think)	Local	off	0.413	2.475

† **DeepSeek V4 Pro** is on a promotional rate of $0.435 in / $0.87 out per million tokens through 31 May 2026. The post-promotion list price is $1.74 / $3.48 (4× higher). All run costs in this paper were billed at the promotional rate. Pricing for every model in this benchmark was last verified on 1 May 2026; treat figures older than ~30 days as stale.

Models we tried and dropped: Google Gemini 3.1 Pro Preview (the endpoint mandates reasoning and was incompatible with the experiment's reasoning-disabled invariant) and Moonshot AI Kimi K2.6 (returned empty content with reasoning off and timed out beyond

180 seconds with reasoning on under strict `json_schema`). These exclusions are statements about API surface, not capability.

Generation procedure

Each model received the same byte-identical system prompt — sourced directly from the K cha khabar production summariser code — at `temperature = 0.2`, `max_tokens = 1536` for reasoning-OFF models and `8192` for the two reasoning-ON variants. Most models routed through OpenRouter; DeepSeek and Qwen 3.6 Max routed direct to their native APIs because OpenRouter's reasoning controls were not honoured by upstream providers in our testing. The system prompt was SHA-256 hashed and the prefix recorded in the results metadata so any drift between runs is mechanically detectable. Two independent generation runs were executed — test-2 (April 27) with eleven models and test-3 (April 30) with the full thirteen — on the same 107 questions.

Rubric

Three axes, integer 1–10 each. The rubric was reverse-engineered from the heuristics actually applied across the 107 questions in test-2, then minimally edited for test-3 to a "lenient ceiling" variant: a 9 may be awarded for "comprehensive, accurate, clean" output without requiring a non-obvious factual capture. In practice the lenient-ceiling rubric ran *stricter* on the 7–9 boundary than the original (mean axis 7.29 vs 7.56) — the grader simply held a higher line for 8 even when 9 was technically more reachable. Detailed anchor tables for all three axes are reproduced in Appendix B.

Throughout this paper, axis scores are presented on the more familiar **/100 scale** (axis × 10), and the overall quality score is the mean of the three axes — also on /100. So a paper that reports "Sonnet 4.6 at 81/100" is reporting an axis-mean of 8.13 on the original 1–10 scale.

Judge

All grading was done by **Claude Opus 4.7 with extended thinking**, single pass per run. The judge sees an anonymised version of each question — option letters only, no model identifiers, no transport metadata — and returns per-option scores plus per-question best-pick with rationale. A separate 20-question informal cross-judge spot-check was run with OpenAI GPT-5.5 and Perplexity Sonar Reasoning. The spot-check established that coarse claims (top model, sign of think/no-think delta, bucket aggregates) are not Anthropic-stylistic-preference artefacts; fine-grained rank claims (rank #4 vs #5) should not be load-bearing on the basis of this work.

§ 05

RESULTS

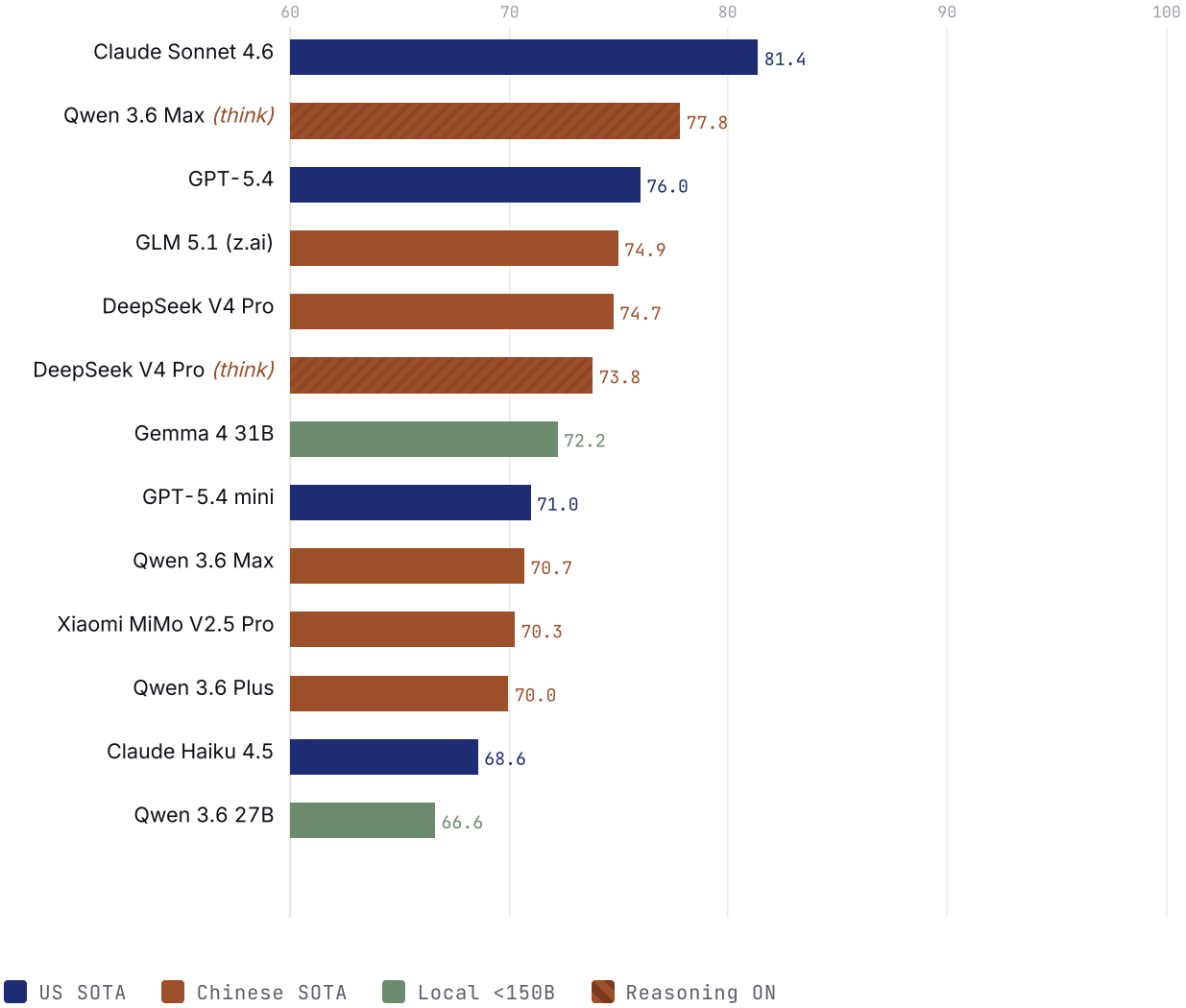
05

Leaderboard

Sonnet at the top, Qwen Max think second, GPT-5.4 third. Open-weight Gemma punches above its bucket. Reasoning helps for one model and hurts for another. The middle of the table reshuffles meaningfully between runs.

FIG 01 · OVERALL QUALITY SCORE, PRIMARY RUN N = 107 PER MODEL

Mean of three axis scores, scaled to /100. Bars coloured by bucket; reasoning-ON variants drawn with hatched fill.



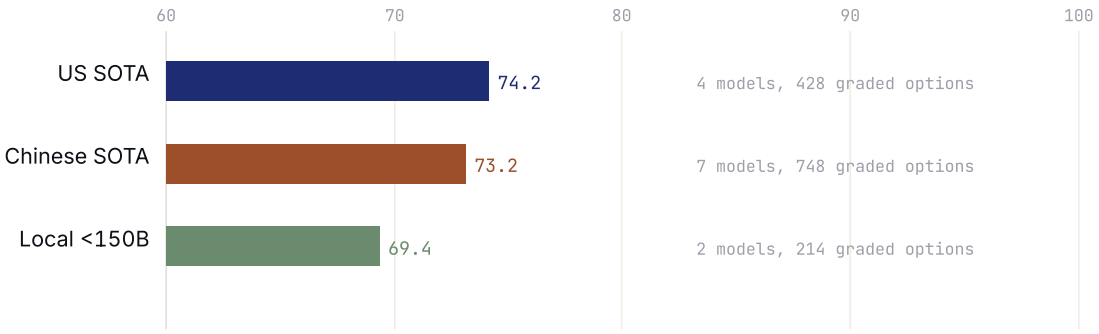
Mean axis quality on a 0–100 scale derived from the three rubric axes (NE prose, EN prose, Topics) at 1–10 each. Bars start at 60 to make the spread legible — the field-wide minimum is 66.6.

#	MODEL	BUCKET	REASONING	NE /100	EN /100	TPC /100	OVERALL /100	#1 FINISHES
1	**Claude Sonnet 4.6**	US	off	85.0	83.0	76.2	81.4	37
2	Qwen 3.6 Max (think)	Cn	on	80.5	78.4	74.4	77.8	17
3	GPT-5.4	US	off	78.9	77.9	71.1	76.0	18
4	GLM 5.1 (z.ai)	Cn	off	77.9	76.2	70.7	74.9	6
5	DeepSeek V4 Pro	Cn	off	77.9	77.7	68.5	74.7	6
6	DeepSeek V4 Pro (think)	Cn	on	76.8	75.5	69.1	73.8	5
7	Gemma 4 31B (it)	Lo	off	74.5	73.6	68.5	72.2	0
8	GPT-5.4 mini	US	off	75.3	72.5	65.0	71.0	5
9	Qwen 3.6 Max	Cn	off	74.6	72.9	64.7	70.7	4
10	Xiaomi MiMo V2.5 Pro	Cn	off	73.7	71.7	65.6	70.3	3
11	Qwen 3.6 Plus	Cn	off	73.1	71.3	65.5	70.0	0
12	Claude Haiku 4.5	US	off	70.9	69.0	65.9	68.6	6

#	MODEL	BUCKET	REASONING	NE /100	EN /100	TPC /100	OVERALL /100	#1 FINISHES
13	Qwen 3.6 27B	Lo	off	70.0	67.4	62.5	66.6	0

FIG 02 · BUCKET AGGREGATES — AVERAGE ACROSS ALL MODELS IN EACH BUCKET

The US SOTA / Chinese SOTA gap closed to 0.32 axis points in test-3 — the within-bucket spread is wider than between-bucket.



§ 06

FINDINGS

Five things this benchmark says

06

i

Sonnet wins both runs, decisively

Claude Sonnet 4.6 is rank #1 in test-2 (ahead by 0.74 axis points) and rank #1 in test-3 (ahead by 1.08 axis points). It wins every axis — Nepali prose, English prose, and topic coverage — in both lineups. The lead compresses when the rubric tightens, but it does not invert. For a workload that prioritises quality and can afford the per-call cost, this is an unambiguous recommendation.

Across the eight failure-mode dimensions tracked in the grader-comment regex (missing facts, hallucination, transliteration, name-spelling slip, party flip, date error, internal NE/EN contradiction, encoding glitch), Sonnet leads or ties on every category except *missing facts*, where every model has a long tail of complaints — comprehensive consolidation under a 1,536-token budget is hard for everyone.

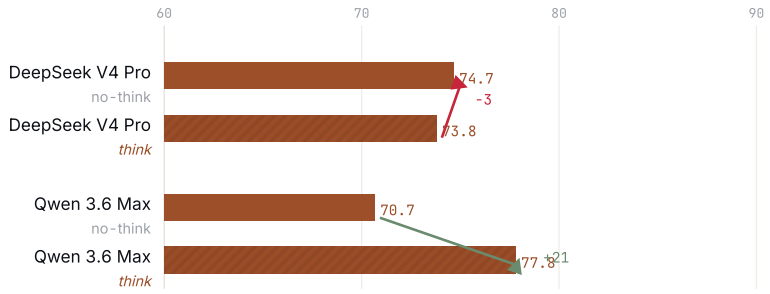
ii

Test-time reasoning has opposite signs across two models

The two direct ablations in this work — DeepSeek V4 Pro think vs no-think, and Qwen 3.6 Max think vs no-think — are the cleanest comparisons we are aware of for Nepali-on-this-task. Same upstream model, same prompt, same temperature; the only invariants changed are the reasoning toggle and the output-token budget.

FIG 03 · REASONING ABLATION THINK VS NO-THINK, PAIRED

Qwen 3.6 Max think gains 21 points; DeepSeek V4 Pro think loses 3.



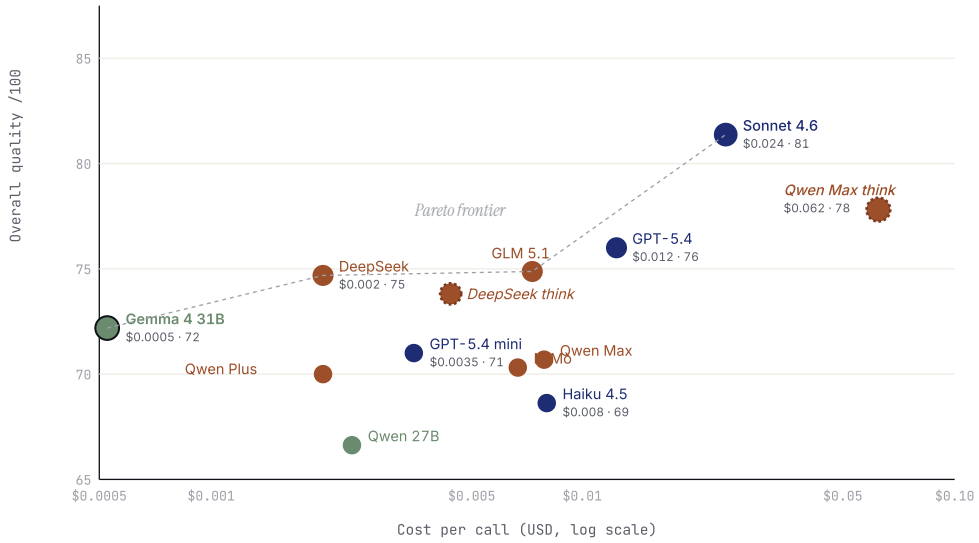
Two interpretations are consistent with the data. One: *different reasoning architectures bring different priors to bilingual generation*. Qwen's reasoning trace appears to help with cross-script consolidation specifically — its Nepali axis jumps from 74.6 to 80.5, English from 72.9 to 78.4. DeepSeek's reasoning trace doesn't change the language axes much (Nepali 77.9 → 76.8, English 77.7 → 75.5). Two: *topics — the entity list — is the dimension that benefits most consistently from thinking* for both models (DeepSeek topics +0.6 of /100, Qwen Max topics +9.7 of /100), implying the trace helps with entity recall under multi-document constraints, but not with prose quality on either side.

For a low-latency production deployment, think variants are not justified by these results unless the application strongly weights the entity list. For an offline batch deployment, Qwen 3.6 Max (think) is competitive with US SOTA at one-third the input cost.

The cost-quality frontier is not where you'd guess

FIG 04 · COST-QUALITY FRONTIER — PER-CALL COST (LOG SCALE) VS OVERALL QUALITY

Gemma 4 31B at $0.0005/call sits well left of every proprietary model. Qwen 3.6 Max (think) buys higher quality but costs more than Sonnet.



The Pareto frontier — the set of models that are not dominated by any other in cost-quality terms — runs through four points. **Gemma 4 31B** at $0.0005/call sits far to the cost-efficient left of every proprietary model and beats GPT-5.4 mini and Haiku 4.5 outright. **DeepSeek V4 Pro (no-think)** at $0.0020/call is the second frontier point. **GLM 5.1** at $0.0073/call sits marginally above DeepSeek (74.9 vs 74.7) at 3.6× the cost — borderline frontier given the rubric's noise floor. **Claude Sonnet 4.6** at $0.0242/call defines the high-quality endpoint. Everything else sits inside the frontier — equal or worse on both axes than something already on the line.

Two practical implications. First: the local-runnable bucket is no longer obviously last. For a Nepali news consolidation deployment specifically, the recommendation is to A/B test Gemma 4 31B against your current proprietary baseline before assuming the proprietary model is required. Second: think variants do not sit on the cost-quality frontier under default budgets. Qwen 3.6 Max (think) reaches 77.8 at $0.062/call — 2.6× more expensive than Sonnet

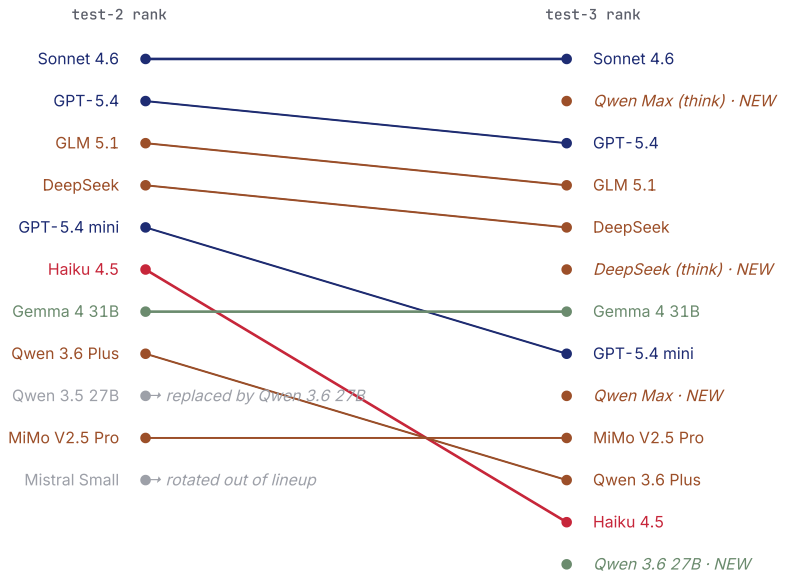
4.6 and ten times slower, with lower quality. Useful for offline batch in a Chinese-API-only deployment context; not for general-purpose substitution.

iv

Top ranks are durable, mid-table is not

FIG 05 · RUN-TO-RUN RANK STABILITY — SAME 107 QUESTIONS, SAME PRIMARY JUDGE, TWO INDEPENDENT RUNS

Sonnet stays at #1; the top-3 set is stable; the bottom of the table reshuffles meaningfully.



Every model dropped 0.5 to 2.0 points between runs because the lenient-ceiling rubric ran stricter on the 7–9 boundary in practice — the grader was simply more conservative on awarding 9 even with permission to. The drop is not uniform: Haiku 4.5, GPT-5.4 mini, and GPT-5.4 dropped most (−13 to −20 on /100), suggesting their test-2 scores were inflated by the more permissive 9 anchor. Sonnet 4.6, DeepSeek (no-think), MiMo, and Gemma dropped least (≤7 on /100), suggesting their test-2 scores were already conservative against the rubric.

The implication for downstream consumers is concrete: *treat absolute Total scores as run-specific and ranks as run-portable for the top half. Bottom-half ranks are more sensitive to which*

models the lineup includes. Claims like "model X is better than model Y" should require a margin of at least 1 axis point on /10 (≥10 on /100), ideally backed by re-run agreement.

V

What broken reasoning actually looks like

DeepSeek V4 Pro think loses to its no-think counterpart on average — but at the per-question level it's closer to a wash (think wins 44, loses 42, ties 21 of 107). The negative average is driven by a long tail of severe regressions, the worst of them −13 on a single question's Total. Inspecting the verbatim reasoning content for the worst regressions reveals a consistent pattern.

Cluster 7314 — police sweep operations in Kaski and Rupandehi. Five outlets reported separate arrest counts on different nights in different districts. The think variant's reasoning trace combined them into a fabricated grand total — *"122+61+61 = 244 in Kaski over three nights. Then Rupandehi on April 25 night: 103. So total 347."* — and the final summary committed to "over 340 arrests". The sources never made that addition. The no-think variant correctly reported "122 + 103 = 225" from two operations, citing the Bikram Sambat dates verbatim from the original copy.

The qualitative pattern across the five worst regressions is the same: **speculative arithmetic on partial evidence**, often combined with date-conversion drift (cycling through several BS↔Gregorian conversions and committing to the wrong one) or late-emerging NE/EN contradictions where a Nepali summary is produced first, then re-derived as English with a different number or date. All three failure modes look like the same underlying pathology: *the reasoning trace treats consolidation as an inference problem (synthesise a unified picture from partial reports) rather than a quotation problem (commit to what at least one source explicitly said).*

For a multi-document news task where outlets disagree on numbers and dates, the inference framing is harmful. The no-think model's terser, more conservative behaviour is closer to wire-copy norms — and to what a working journalist would do. Hedge phrase counts in the bad-case reasoning traces ("maybe", "perhaps", "or around", "I think") correlate with regression severity.

The trace is *aware* of its uncertainty; it just doesn't let that uncertainty propagate into the final answer.

For DeepSeek V4 Pro specifically, the recommendation from this benchmark is to run it with thinking disabled for offline batch — and to accept that production reliability (where thinking-on still wins by virtue of fewer empty-content failures) is a separate axis not captured here.

§ 07

PER-CALL PERFORMANCE

07

Latency, tokens, and dollar cost

Test-3 retained full per-call performance metadata for every successful call: latency, prompt and completion tokens, separate reasoning tokens for think variants, and forensic per-call dollar cost computed from each model's published rates as of 1 May 2026. The aggregate run cost across 1,390 successful calls was **≈$15.40** at corrected rates; mean call latency was **27 seconds** across the lineup, but the within-lineup spread is two orders of magnitude.

MODEL	MEAN LATENCY	IN TOKENS	OUT TOKENS	REASONING TOKENS	COST PER CALL	RUN COST
Qwen 3.6 Max (think)	106 s	3,175	4,022	3,445	$0.0624	$6.67
Claude Sonnet 4.6	15 s	4,647	687	—	$0.0242	$2.59
GPT-5.4	8 s	2,568	390	—	$0.0123	$1.31
Claude Haiku 4.5	8 s	4,646	673	—	$0.0080	$0.86
Qwen 3.6 Max (no-think)	11 s	3,177	487	—	$0.0079	$0.85
GLM 5.1 (z.ai)	30 s	4,664	700	—	$0.0073	$0.79
Xiaomi MiMo V2.5 Pro	10 s	4,845	634	—	$0.0067	$0.72
DeepSeek V4 Pro (think) †	95 s	3,580	3,240	2,711	$0.0044	$0.47
GPT-5.4 mini	4 s	2,568	355	—	$0.0035	$0.38
Qwen 3.6 27B	9 s	3,177	450	—	$0.0024	$0.26
DeepSeek V4 Pro (no-think) †	16 s	3,580	524	—	$0.0020	$0.22
Qwen 3.6 Plus	10 s	3,177	487	—	$0.0020	$0.21
Gemma 4 31B (it)	28 s	2,555	393	—	**$0.0005**	**$0.05**

† DeepSeek V4 Pro per-call costs reflect the promotional rate ($0.435 in / $0.87 out per million tokens) in effect through 31 May 2026. At post-promo list prices ($1.74 / $3.48), DeepSeek V4 Pro (think) per-call rises to ≈$0.0175 and (no-think) to ≈$0.0080.

Pricing volatility note. An earlier version of this paper used $/M rates of 0.40 in / 4.00 out for Qwen 3.6 Max and 0.195 in / 1.56 out for Qwen 3.6 27B. Both were incorrect: the first appears to have been copied from an older Qwen tier; the second is the OpenRouter Qwen 3.5 27B price, not Qwen 3.6 27B's DashScope rate. The corrected DashScope rates are $1.30 / $7.80 for Qwen 3.6 Max and $0.413 / $2.475 for Qwen 3.6 27B (verified against DashScope's pricing page on 1 May 2026). All cost figures and the cost-quality scatter in this paper reflect the corrected rates.

A few observations worth pulling forward. The think variants spend most of their wall-clock time in reasoning rather than answer generation: DeepSeek's think trace burns 84% of its 95-second budget in the trace; Qwen 3.6 Max think burns 86% of 106 seconds. For a real-time-facing application that must finish well under a minute, neither is viable — the user would have left the page before the answer arrived.

The latency gap between Sonnet (15 s) and Haiku (8 s) is smaller than one might expect; the gap between GPT-5.4 mini (4 s) and GPT-5.4 (8 s) is what the OpenAI naming scheme advertises. GLM 5.1 at 30 seconds is unusual — its average latency is in the same range as Gemma 4 31B (28 s), but its prompt size and output size suggest a slower upstream provider rather than longer inference. Latency in this benchmark is necessarily a property of the *route*, not the *model*.

Devanagari tokenisation disparity also matters. The same Nepali summary length produces different token counts across model families — different tokenisers split Devanagari differently. The MiMo and Qwen tokenisers are notably more verbose on Devanagari than GPT-5 family or Gemma. Per-call cost comparisons are *not* adjusted for this; readers building cost models for a particular language should account for it directly.

§ 08

LIMITATIONS

08

What this benchmark cannot say

This is a single-author study with N=107 cluster-questions, two generation runs, and one primary judge. It is not peer reviewed. The known holes, in roughly decreasing order of severity:

1. **Single primary judge.** All 107 × 2 questions were graded by Claude Opus 4.7. The 20-question informal cross-judge spot-check with GPT-5.5 and Perplexity Sonar Reasoning is a sanity check on coarse claims, not a Cohen's-κ-grade calibration. A formal 3-judge × 3-run design is committed for v0.3.
2. **No human gold labels.** The benchmark is fully LLM-graded. Even N=1 human-expert grading on a 20-question sample would meaningfully harden it; that work has not been done.
3. **N=107 is small for stable per-axis claims.** Per-model averages within ±0.5 axis points (±5 on /100) of each other are not statistically distinguishable under any reasonable test. Standard errors are not reported.
4. **Single sample per (model, question).** Outputs were generated at `temperature=0.2` once per cell. The variance attributable to sampling is unmeasured — though it can now be partially inferred from the test-2 vs test-3 deltas at the model level.
5. **LLM-judge bias.** The judge (Anthropic Opus 4.7) graded outputs that include 2 Anthropic models (Sonnet 4.6, Haiku 4.5). It is plausible that the judge has a stylistic preference for outputs that resemble its own family's prose. The cross-judge spot-check addresses this only at coarse granularity.
6. **Rubric is descriptive, not pre-registered.** It was reverse-engineered from heuristics applied during test-2 grading and minimally edited for test-3. Future versions should pre-register the rubric.
7. **Source articles are RSS excerpts, not full text.** The grader's "ground truth" is itself thin in the long tail of facts. Some "missing facts" complaints may be against details that weren't in the excerpts the grader saw either.
8. **Same-lab clustering.** Anthropic ×2, OpenAI ×2, Qwen ×4 in the test-3 lineup. Bucket averages reflect lab strategy as much as model ability.
9. **API-surface exclusions.** Gemini 3.1 Pro and Kimi K2.6 dropped for incompatibility with the reasoning-disabled invariant under the chosen response-format settings — capability claims are about the API surface, not the underlying models.

10. **Devanagari tokeniser disparity.** Per-call cost numbers are not adjusted for the fact that different tokenisers split Devanagari differently — some models emit ~30% more tokens than others for the same Nepali summary.
11. **Reliability is not in the leaderboard.** Empty content, schema violations, time-outs are tracked in the per-call metadata for test-3 but not aggregated into the score. A reliability-aware leaderboard is committed for v0.4.
12. **Test-3 grader rubric drift ("lenient-ceiling v1").** The rubric was edited between runs to permit a 9 for "comprehensive, accurate, clean" prose without requiring a non-obvious detail. The lenient ceiling did not produce score inflation in practice (mean axis dropped 7.56 → 7.29), but the run-to-run deltas are confounded with this rubric change.

This is a list of things we know are wrong, not a list of things we think are unimportant. Most can be fixed; some need a budget the project does not yet have.

§ 09

FUTURE WORK

What's next

09

V0.3 · CALIBRATION

Formal 3-judge × 3-run × 20-question calibration with Cohen's κ on best-pick and Spearman ρ per axis. Anthropic Opus 4.7, OpenAI GPT-5.5, Google Gemini 3.1 Pro. Tooling already in place; remaining work is the small spend.

V0.3 · CROSS-VALIDATE

Score the thirteen models against the IndicGenBench Nepali split (~500 single-article items, ROUGE/chrF). Direct comparison with prior work clarifies whether the rubric-graded leaderboard correlates with traditional metrics or diverges meaningfully.

V0.3 · HUMAN GOLD

Add 1 human-expert sample (N=1, 20 questions) if a Nepali journalist is willing to grade. Anchors the LLM-judge results to ground truth.

V0.4 · RELIABILITY AXIS

Combine rubric quality with empty-content, schema-violation, and timeout rates from per-call metadata into a single deployability score. The current leaderboard rewards quality on successful outputs; production cares about the success rate too.

V0.4 · DIALECT RUBRIC

Re-grade with a stricter rubric that distinguishes between Nepali speakers' regional dialect preferences (Kathmandu vs Eastern vs Western Nepal). The current rubric implicitly encodes the grader's assumptions about "good Devanagari prose."

V0.5 · REASONING-ON AT SCALE

Extend the reasoning ablation beyond the two models tested directly to GPT-5.4, Sonnet 4.6 in extended-thinking mode, GLM 5.1 with reasoning, and Gemini 3.1 Pro. Quantify whether the DeepSeek pattern (think hurts) or the Qwen pattern (think helps) generalises.

§ 10

ACKNOWLEDGEMENTS

10

The journalists who wrote the corpus

This benchmark redistributes **headlines and ≤280-character excerpts** from 1,310 articles published by 23 Nepali news outlets between February and April 2026. Original copyright belongs to each publisher. Excerpts are redistributed under fair-use principles for non-commercial research only; takedown is unconditional.

Khabarhub (Nepali)	188	Gorkhapatra Online	152	Nagarik News	148
OnlineKhabar (Nepali)	129	Nepal Press	99	Himal Press	98
News of Nepal	78	Rajdhani National Daily	60	Setopati	56
Khabarhub (English)	42	Lokaantar	42	The Rising Nepal	36
Ratopati	34	Nepal News (English)	34	Baahrakhari	33
Thaha Khabar	26	myRepublica	15	Nepal Lead	13
Nepal Samaya	8	BBC News Nepali	5	The Himalayan Times	5
The Annapurna Express	5	OnlineKhabar (English)	4		

> *Without the daily reporting of the journalists at these outlets — the people who attend press conferences, work the sources, and file copy on deadline in two languages — the benchmark has no substrate. They are the unpaid co-authors of every model output evaluated here.*

Methodologically, the work also leans on three external bodies of research: the **NLUE** Nepali NLU benchmark (Singh et al., 2024) for the survey of what does and does not exist in Nepali NLP today; the **Prometheus-2** and **RubricEval** rubric-based judging methodology canon; and

the cautions raised in *"How Reliable is Multilingual LLM-as-a-Judge?"* (Findings EMNLP 2025) on per-language Kappa instability — the cross-judge spot-check is essentially that paper's recommendation, scoped down for cost.

The Nepali NLP community at IRIIS Nepal, Kathmandu University ILPRL, Tribhuvan University, NAAMII, and a long roll-call of independent contributors maintain the pretrained models, datasets, and benchmarks (NLUE especially) that make Nepali NLP a phrase that means something in 2026. This work would not exist without theirs.

§ 11

ABOUT & ACCESS

Who we are. How to reach us. How to get the data.

Outback Yak

Outback Yak is an Australian-incorporated AI & engineering studio (ABN 11 672 730 773) based in Melbourne. We build AI agents, cloud-and-automation systems, and apps & platforms for clients across Australia and the wider region. K cha khabar is one of our products — a research-grade Nepali news intelligence platform, deliberately operated outside Nepal to insulate the Nepali diaspora's information environment from in-country pressures on the press.

K cha khabar

K cha khabar (kchakhabar.com) is a cross-publisher Nepali news intelligence platform: AI-clustered stories with bilingual summaries, ownership transparency across thirty-plus publishers, trending topics, breaking news. Every headline links to its source. The platform is built for the Nepali diaspora, for engaged in-country readers, and for researchers of the Nepal media ecosystem. NepNewsCluster is the formal evaluation backing the benchmark choices made in K cha khabar's production pipeline.

Data & code access

The accompanying data and code repository is currently private. We will share access — including the 107 anonymised cluster questions, the model→option answer key, the full grader output, and all reproduction code — with researchers, journalists, and engineers who can describe their use case briefly. Send a one-paragraph note to research@kchakhabar.com describing what you would do with the data, and we will respond within a few business days. There is no fee.

Citation

```
@misc{nepnewscluster2026,
  author       = {Outback Yak},
  title        = {NepNewsCluster v0.2: A bilingual Nepali-English
                  news consolidation benchmark across 13 frontier
                  and open-weight LLMs},
  year         = {2026},
  howpublished = {Outback Yak Research},
  note         = {v0.2 — primary judge (Opus 4.7), N=107,
                  two generation runs, 13 models in primary lineup}
}
```

APP. A

GENERATION PROCEDURE

Self-contained methodology reference

Sampling and decoding

- **Temperature = 0.2.** Keeps decoding ~95% deterministic but lets the model break out of degenerate states. Lower than the popular default of 0.7 because the task is factuality-driven and does not benefit from stylistic variance; higher than 0.0 because some models (notably Haiku and DeepSeek no-think) emit subtly worse JSON at strict greedy decoding.
- **max_tokens = 1536** for reasoning-OFF models, **8192** for reasoning-ON variants. The 1,536 budget covers the bilingual headline + summary + entity list + slug comfortably; the 8,192 budget covers reasoning trace plus answer.
- **response_format = json_schema** with `strict: true` for ten of thirteen models. The three native-API exceptions (DeepSeek both variants, Qwen 3.6 Max both variants, Qwen 3.6 27B) use `json_object` because their native APIs do not support strict schemas as of April 2026.
- **Concurrency:** 13 models per cluster fanned out via parallel calls; one cluster at a time. Atomic JSON write after every call → fully resumable on crash.
- **Per-call retry:** up to 3 attempts on transient errors (HTTP 5xx/429, timeouts, network). No retry on 4xx or empty-content failures — those are surfaced as model failures rather than retried into success.
- **Request timeout:** 240 seconds — DeepSeek and Qwen 3.6 Max think variants can take 60–110 seconds.

Prompt structure

The system prompt establishes the role ("bilingual news editor for a Nepali aggregator"), the rules (neutrality, headline length caps, no transliteration, content-flag handling for casualty/minor stories), and the entity-extraction contract. The user prompt enumerates each source article in a structured block:

```
[1] News of Nepal (en) — 2026-04-24 04:52 UTC
  Tags: news_event, politics
  EN: Speaker Devraj Ghimire calls all-party meeting...
  NE: सभामुख देवराज घिमिरेले सबै दलको बैठक बोलाए
  Excerpt: The Speaker has called all parliamentary parties...
```

```
[2] OnlineKhabar (ne) — 2026-04-24 06:13 UTC
  ...
```

The K cha khabar production summary is *not* in the prompt. Each model is asked to produce its own consolidation from the article-level inputs only.

Sampling distribution

The 107 clusters span themes including national politics, district-level governance, foreign policy, sports (cricket, football), business and economy, education policy, religious events, weather and disaster reporting, road accidents, and obituaries. The dominant theme is politics (40-odd clusters), reflecting the bias of Nepali RSS feeds in the April 2026 sample window. Future work should explicitly down-sample politics to surface non-political performance differences.

APP. B

RUBRIC (LENIENT-CEILING V1)

Three axes, integer 1–10 each

The full anchor table for each axis. Total = NE + EN + Topics, range 3–30, presented in the body of the paper as a /100 axis-mean.

Axis 1 — Nepali (NE) prose quality

Judges `headline_ne` + `summary_ne` together.

SCORE	ANCHOR
9–10	Captures a unique factual detail others miss, or is otherwise outstanding. Clean prose. All WHO/WHAT/WHERE/WHEN threads woven naturally. *Lenient mode*: also award 9 if the work is comprehensive, accurate, and clean even without a unique detail.
8	Comprehensive and accurate. All major facts present. No errors. Reads like clean wire copy.
7	Adequate but limited. Concise, missing some context. No factual errors, just under-specified.
6	One small error: name spelling slip, transliteration variant the source doesn't support, awkward phrasing that drops nuance, generic title where source named the person.
5	Disqualifying-class error: encoding glitch, wrong title or role, internal contradiction with the EN side, or BS↔Gregorian date conversion off by a month.
≤ 4	Hallucinated proper nouns, fabricated quotes/numbers, cause-of-death errors, multiple compounding errors.

Axis 2 — English (EN) prose quality

Same scale as NE, applied to `headline_en` + `summary_en`. Common cross-axis failure modes:

- **NE correct, EN wrong**: translation that flips meaning (लेक "highland pasture" → "Lake"; सम्पत्ति शुद्धीकरण "money laundering" → "Asset Recovery"), party flip (Oli labelled "Nepali Congress President").

- **NE wrong, EN correct**: Devanagari date conversion mistake (Vaishakh 11 → "अप्रिल ११" instead of "अप्रिल २४"), title slip (अध्यक्ष for सभामुख).
- **Both wrong, in different ways**: internal NE/EN contradiction. Penalised on both axes; this is worse than a single error in either.

Specific EN deductions: −2 to −3 for translation that flips meaning; −2 for code-switching (Devanagari word inside EN prose); −1 to −2 for systematic BS-to-Gregorian date errors; −1 for plausible name transliteration drift, −2 if it changes the person.

Axis 3 — Topics (entity list) quality

Judges the `entities` list.

SCORE	ANCHOR
9–10	Comprehensive coverage, all correctly typed, AND a non-obvious entity others missed. *Lenient mode*: also award 9 for comprehensive + correctly typed even without a non-obvious entity.
8	Most entities, all correctly typed, captures most named persons/places/orgs.
7	Mostly correct, one minor type mistake or one missing obvious entity.
6	Missing some entities, one clear type error.
5	Multiple type errors or fewer than 2–3 entities when sources offered 6+.
≤ 4	Fabricated entity in the list, or majority of entities mistyped.

Frequently-penalised entity-type sins

- Species or animal classified as `place`
- Day-of-week classified as `event`
- Hospital or airport classified as `org` instead of `place`
- Religion or ethnicity classified as `org`
- Deity name classified as `event`

APP. C

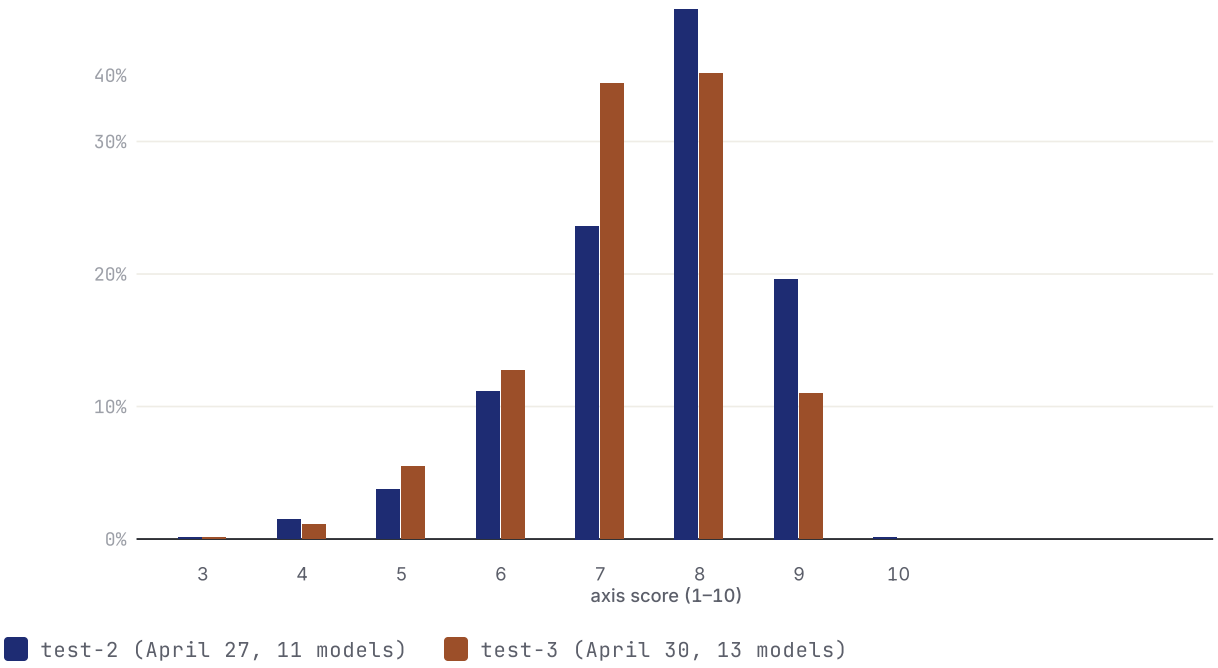
SCORE DISTRIBUTION

How the grader actually used the scale

Across both runs the grader (Claude Opus 4.7) used most of the 1–10 range, but the modal score is 8 in both. Test-3's lenient-ceiling rubric did not produce more 9s — it produced fewer (10.97% vs 19.66%), more 7s (34.34% vs 23.57%), and zero 10s (vs 6 in test-2). The grader interpreted "lenient ceiling" as "be more conservative on the 7→8→9 boundary" rather than "give out more 9s".

FIG 06 · AXIS-SCORE DISTRIBUTION — PER-AXIS 1–10 INTEGER SCORES, BOTH RUNS

Mean axis dropped 7.56 → 7.29 between runs; the 9-share collapsed from 20% to 11%.



Distribution of integer axis scores across all (option × axis) cells. The grader treated 8 as the default ceiling and 9 as scarce in both runs; the lenient-ceiling rubric did not change that.

One implication for downstream consumers building tooling on top of this kind of grader: *expect a Gaussian-ish spread centred on 8*, not a uniform distribution across 1–10. Any score difference of less than 0.3 axis points (3 on /100) at the model level is operating well within the noise floor of how this rubric is actually applied.

Outback Yak is an Australian-incorporated AI & engineering studio (ABN 11 672 730 773) based in Melbourne. We build AI agents, cloud-and-automation systems, and apps & platforms across Australia and the wider region.

K cha khabar — research-grade Nepali news intelligence — is one of our products.

OUTBACK YAK

outbackyak.io

hello@outbackyak.io

Melbourne, Australia

K CHA KHABAR

kchakhabar.com

research@kchakhabar.com

For data & reproduction code access