

Obligatory Slots: Under Reference-Free Evaluation, Dropping a Distinction the Source Never Made Is Free

Luxshan Thavarasa¹  Jubeerathan Thevakumar²  Sivasuthan Sukumar² 

¹Independent Researcher, Colombo, Sri Lanka

²Department of Computer Science and Engineering,
University of Moratuwa, Colombo, Sri Lanka

luxshan.20@cse.mrt.ac.lk

jubeerathan.25@cse.mrt.ac.lk, sivasuthan.26@cse.mrt.ac.lk

Abstract

When the target language must mark a distinction the source does not, a translation system must still pick a value. English *they came* becomes one Tamil verb if those who came are people and another if they are not; English does not say which. We ask what automatic evaluation charges when the choice is wrong, or never made. We release TamilLing-Bench,¹ an English→Tamil challenge set for three verb-agreement distinctions, checked by a morphological analyser: rationality (திறனை *tiṇai*), gender, and number as a control. We pre-registered a prediction that neural metrics would not notice a single wrong morpheme. It is wrong for the neural reference-based metrics: COMET-22 charges $8.52\times$ the score difference that human raters reliably notice. On the systems' own errors, both reference-free metrics rank the wrong form first more often than not, and a legitimate Tamil form that leaves the distinction out is not charged at all. **Under reference-free evaluation, dropping the distinction is free**; reference-based metrics do charge for it. xCOMET finds the error's region but not the morpheme. Inside one model, activation patching finds the same asymmetry: a distinction English marks is settled by the middle of the network, one it leaves unmarked only near the output.

1 Introduction

English *they came* has two Tamil translations. If the referents are people, the verb is வந்தார்கள் *vantārkaḷ*; if they are not, it is வந்தன *vantaṇa*. The English sentence never chose, and Tamil cannot avoid choosing, because the finite verb agrees with its subject in rationality (திறனை *tiṇai*), gender and number. We call such a slot *obligatory*: every well-formed target realisation commits to a value the source does not determine. The stake is

¹Code, data and system outputs: <https://github.com/Luxshan2000/TamilLingBench>

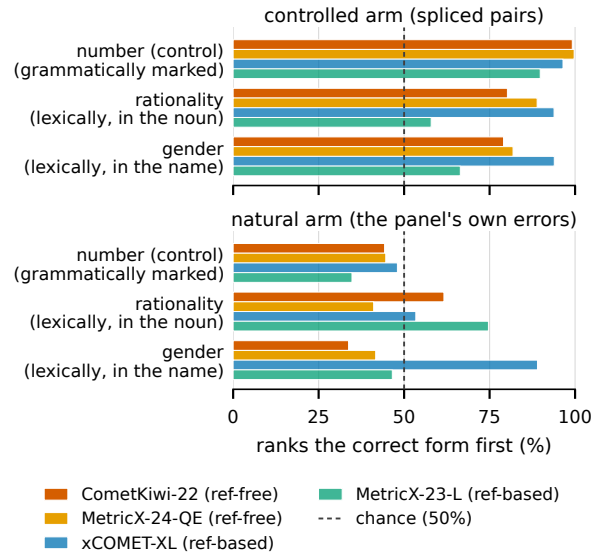


Figure 1: **What each metric can rank, ordered by what the English source makes recoverable.** Each bar is the share of pairs on which a metric ranks the correct member first; 50% is chance. Pairs differ in one grapheme-cluster span, the analyser confirming the rejected member carries the wrong value. **Top:** constructed pairs. The two reference-based metrics with headroom disagree with each other here, so this ordering is **not** a reference-free phenomenon. **Bottom:** the panel's own errors, where the reference-free/reference-based split *is* clean, the result §5 defends. COMET-22 and chrF2 are omitted: they sit at 100.0% everywhere, partly definitionally, since here the correct variant *is* the reference.

not stylistic. அகறினை *akriṇai* agreement, the class Tamil uses for everything that is not a person or a deity, on a human referent is a dehumanising error. This paper asks what automatic evaluation charges when the value is wrong, and when the system quietly declines to pick one at all. The answer, on a morphologically checked English→Tamil benchmark we release, splits by metric type. Reference-based metrics charge a large penalty, falsifying our own pre-registered prediction. Reference-free metrics do not: on the er-

rors the systems themselves make both rank the wrong output first more often than not, and the legitimate Tamil form that drops the distinction costs nothing they can distinguish from zero.

Tamil marks several such distinctions on the finite verb, and a purpose-built, morphologically checked benchmark is necessary rather than nice-to-have because the standard annotation infrastructure cannot express most of them. UD_Tamil-TTB contains 0 Gender=Fem tokens and 0 -ஆள் *-āṇ/-āḷ* finite verbs, so a tagger trained on it cannot express the masculine/feminine contrast at all; Tamil is absent from UniMorph entirely; and in total the annotation cannot express 7 categories a reference grammar treats as central (aspect, causation, clusivity, deixis, evidentiality, honorific contrast, modality). Appendix D and Table 11 give every count, each a direct count over the released treebanks. These are reproducible counts, not arguments about anyone’s competence.

Where this sits. Challenge sets in the contrastive tradition, from ContraPro (Müller et al., 2018) to Morpheval (Burlot et al., 2018), test whether a system *propagates* a feature present in the source. CoCoA-MT and the IWSLT formality tasks (Nadejde et al., 2022; Anastasopoulos et al., 2022) meet the unmarked case but run it as an uncontrolled baseline row, and GENDER1PERSON (Popović and Lapshinova-Koltunski, 2025) measures the resulting skew for one slot without asking what evaluation charges for it. The question here is what happens when there is nothing to propagate and a value must still appear, and what a metric can see when it does; §9 draws the full comparison.

The metric question. An obligatory-slot error is one morpheme in an otherwise fluent, adequate sentence, and whether automatic evaluation can see it decides whether the error type is actionable at all. We pre-registered the prediction that neural metrics would not charge a human-noticeable penalty for it. **That prediction is falsified for the reference-based metrics**, and we report the falsification rather than narrowing the claim after the fact: 3 of the 4 neural metrics with a published calibration charge well above the 75%-agreement threshold (Kocmi et al., 2024), and so does chrF2, a character *n*-gram metric the prediction never covered.

Design, and one thing it got wrong. Every item appears in four conditions that vary only where the

disambiguating cue sits. To be plain about what is new: **only the cue-free condition has no answer in the source**. The other three carry a gold value, supplied by the instruction, by the sentence itself, or by the sentence before, and on those the benchmark is contrastive evaluation in the usual sense. The cue-free condition is the one the contrastive tradition runs, at most, as an unscored baseline row (§9), and it is where the forced choice this paper is about actually happens. Our first release had three independent construction defects in one slot, which inverted a headline we had already drawn; §8 reports what broke and how the audit caught it, because the failure mode generalises to any benchmark of this shape.

Contributions. (i) A measurement of what automatic metrics charge for obligatory-slot errors. Our own prediction is falsified for the neural reference-based metrics. The finding that survives, that dropping the distinction is free under reference-free evaluation, is established on the outputs the systems actually produce, with positive controls on the same sentences. (ii) **TamilLingBench**, a generator and an FST-based checker for a linguistically controlled Tamil challenge set, released with per-item judgments and 12 construction guards, verified by two native speakers of Tamil. (iii) Three construction confounds in one slot, with the audit that found them. (iv) A pre-registered causal dissociation inside one model, reported with the pre-registered probe failure from the same frozen document. (v) Provisional model measurements, including the withdrawal of our own first headline, and the annotation negative results above.

2 Obligatory Slots in Tamil

The scored slots are rationality, gender and number as a control; Table 5 in Appendix A gives each one’s Tamil realisation and prior MT coverage. The contrast with source ambiguity separates an obligatory slot from contrastive conditioning (Vamvas and Sennrich, 2021b): there the source admits two readings and the question is which the system selects; here it admits one and is silent on a distinction the target must realise. A system cannot hedge: except where the language supplies a neutral exponent, which is what makes the gender slot a cautionary tale (§8). All three are realised on the finite verb, which makes this one system rather than a checklist. **Rationality** (திணை *tiṇai*)

splits referents into உயர்திணை *uyartinaṭai*, people and deities, against அஃறிணை *akriṇai*, everything else, and agreement follows: வந்தார்கள் *vantārkaḷ* against வந்தன *vantaṇa*. English has no exponent of it, though the subject noun usually makes it lexically recoverable, *the dogs* against *the students* — a different thing from marking it, and what §5 turns on. On a plural அஃறிணை subject the licit exponent is -அன *-ana*, not the singular -அது *-atu*, on the ruling of a native speaker among the authors. **Gender** in the verb is வந்தான் *vantāṇ* / வந்தாள் *vantāl* / வந்தார் *vantār* — masculine, feminine, and an *epicene* form, a licit Tamil word committing to neither. English marks it on a name or a pronoun, so it too is lexically recoverable. **Number** is the control: both languages mark it, so a competent system should be at ceiling.

What is not here, and why. Second-person honorificity and first-person-plural clusivity were scored in earlier releases and are withdrawn as construction defects rather than findings (Appendix A); §5 states what withdrawing them costs its recoverability result. Second-person address in the indicative is **binary** — நீ *nī*/நீர் *nīr* against நீங்கள் *nīṅkaḷ*/தாங்கள் *tāṅkaḷ*, on the ruling of a native speaker among the authors, and the shipped items pooled it with a differently sized imperative paradigm under one chance rate. A fourth slot, deictic proximity, is under construction and is not reported. All items are standard written Tamil, and Literary/Spoken diglossia is *orthogonal* to the choices tested here (Prasanna and Arora, 2024).

3 TamilLingBench

Four conditions over one frame. With the slot inventory fixed, the benchmark varies one thing. Every item is a minimal set over a single frame, and the conditions differ only in where the value can be found. **Instructed:** the instruction names the target Tamil form outright. This is a capability ceiling rather than a translation task: a system that fails here cannot be read as *defaulting* anywhere else. **In-sentence cue:** the English source itself determines the value, and gold exists. **Prior-sentence cue:** the value is recoverable only from the sentence before; gold exists, and following Bawden et al. (2018) the gold values are balanced so a system that ignores context scores exactly chance. **Cue-free:** the source is genuinely silent, **there is no gold and no answer is wrong**, so we report the distribution of emitted values against a Tamil

corpus prior, never against a notion of the correct default, which does not exist. **The comparison we report is in-sentence against cue-free**, which holds the frame and the cue type fixed and deletes the cue. **In-sentence against prior-sentence is not reported for gender**, because there the two conditions cue the value through different channels (§8); §6 reports it for rationality only. Table 1 shows one released item per slot and cue.

Construction. The release contains 2873 items generated from a template grammar with a fixed seed (20260807), split into a public portion and a blind holdout at the *template* level: holding out items is worthless once the generator ships. Validation is at that level too, plus a stratified item audit, because templates decide whether the contrast is real and fillers only vary lexical content. The source↔gold audit of §8 is run on every release of the set. Vamvas and Sennrich (2021b) show that contrastive results generalise only when the hypothesised contrast is independently motivated; ours is a typological obligation. The distributional gap to natural Tamil they also ask for is not computed (Limitations).

Automatic checking. Outputs are scored by a finite-state analyser built on the ThamizhiMorph transducer union (Sarveswaran et al., 2021) plus a hand-written fallback layer, with an ISO 15919 canonical scoring key (Rajan, 2018). We report the ablation rather than one number (Table 10): coverage on the UD_Tamil-TTB test verb forms is 38.6% lexicon-only and 69.3% with fallback rules. **The checker is not 100% accurate and we do not claim it is;** items it cannot analyse are UNPARSED, reported separately, never dropped. Scoring is universal after context filtering, since வந்தார்கள் *vantārkaḷ* is ambiguous between honorific-singular and rational-plural, and the tooling needed four documented repairs at the scoring layer first (Appendix C).

Outcome taxonomy, and what commitment means. Each judgment is CORRECT, WRONG, UNDECIDABLE, one of three AVOIDANT outcomes (no carrier, a value-neutral form, or a negated clause, which in Tamil can strip the agreement), or UNPARSED. A system *commits* on an item when it emits a form that takes a value in the slot — CORRECT or WRONG, but not AVOIDANT and not UNPARSED; the *commitment rate* is the share of items on which it does so, and accuracy is always com-

Slot	Cue	English source	Tamil gold (ISO 15919)	Value
rationality	in-sentence	The dogs came at dawn.	நாய்கள் வந்தன. <i>nāykaḷ vantaṇa.</i>	அஃறிணை <i>akriṇai</i>
	in-sentence	The students came at dawn.	மாணவர்கள் வந்தார்கள். <i>māṇavarkaḷ vantaṛkaḷ.</i>	உயர்திணை <i>uyartiṇai</i>
	cue-free	<i>The sky had been overcast since dawn. They came at dawn.</i>	<i>no gold: any value is licit</i>	none
gender	in-sentence	Akanakai came at noon.	அகநகை வந்தாள். <i>akanakai vantaḷ.</i>	feminine
	in-sentence	Akaran came at noon.	அகரன் வந்தான். <i>akaran vantaṇ.</i>	masculine
	cue-free	<i>Aram, our new neighbour, had been waiting at the station since noon. Aram came at noon.</i>	<i>no gold: any value is licit</i>	none
number (ctrl.)	in-sentence	The students spoke outside the hall.	மாணவர்கள் பேசினார்கள். <i>māṇavarkaḷ pēciṇārkaḷ.</i>	plural
	in-sentence	The student spoke outside the hall.	மாணவன் பேசினான். <i>māṇavaṇ pēciṇāṇ.</i>	singular
	cue-free	<i>A short notice about the theft had gone up. The accused spoke outside the hall.</i>	<i>no gold: any value is licit</i>	none

Table 1: **One released item per slot and cue.** In-sentence items come in minimal pairs whose English differs only in the cue (for gender, the name itself) and whose Tamil verb differs in the agreement morpheme; the cue-free item keeps the frame and deletes the cue, so nothing in the source determines the value and no gold exists. In the cue-free gender item the systems’ dominant output is the epicene வந்தார் *vantaṛ*, a well-formed form that commits to neither gender; §3 scores it as AVOIDANT, never as an error. Cue-free items carry a preceding context sentence (italic) that says nothing about the value. The prior-sentence and instructed conditions use the same frames (§3).

puted among committed items only. **AVOIDANT is never folded into “incorrect”** — the decision that makes §5’s main result visible at all, since a benchmark scoring the epicene form as an error would report an accuracy collapse where there is a refusal. A cell with fewer than 30 committed items or below 30% commitment prints *n/a*.

4 Experimental Setup

The panel is 13 systems, chosen to make three contrasts available rather than to rank anything: Gemma 3 at 4 sizes for a scale series inside one family; Qwen3 at three sizes and Llama-3.1 for family variation at roughly fixed scale; NLLB-200-3.3B and IndicTrans2 as dedicated encoder–decoder systems; and Sarvam-Translate, a Gemma-3-4B derivative fine-tuned for Indic translation, which against Gemma 3 4B varies the fine-tuning objective alone. 11 are prompted and 2 dedicated MT systems, read off the data rather than declared; there is no frontier commercial system, by decision. Scoring is on generated output, not forced choice: $\log P(\text{ref}) > \log P(\text{contrast})$ is unusable on systems that do not expose likelihoods. Decoding is greedy for the prompted systems and beam search at the deployed defaults for the dedicated MT systems, on seed 20260807, over 37 runs on transformers 5.12.1; each run records its re-

solved model revision, config hashes and decoding block. The dedicated MT systems take no prompt at all (Appendix A).

Prompt choice moves these numbers. Across the 73 cells where more than one prompt yields a reportable accuracy, the median spread is 7.5 points and the worst is 82.5. **Every headline cell is therefore a single-prompt measurement**, held fixed across systems so that between-system comparisons are clean, and no absolute accuracy here should be read as a property of a system independent of its prompt. The prior-sentence condition uses the prompt that marks the target sentence and tells the system not to translate the preceding one (P4_context in the released prompt set). Unreadable outputs are UNPARSED with their own denominator, mean 6.5% and worst cell 45.8%.

The cue-free condition has no gold, so we report the emitted distribution against a Tamil corpus prior pooled over 3 corpora, always a range because every slot is fragile: which corpus one picks moves the prior more than how many tokens one counts. We therefore say an output rate *deviates from general Tamil text frequency* and no more; the stronger claim, that it matches a system’s *own* training rate, is unavailable because no evaluated system’s training mixture was measured.

Metric	ref.	scale	win	$\times$ agr.	$\Delta_{av}\%$	CI $\ni$ 0
COMET-22	yes	0–1	100.0	8.52	3.96	no
xCOMET-XL	yes	0–1	62.3	6.55	2.23	no
MetricX-24	yes	0–25	90.2	<i>n/a</i>	7.56	no
MetricX-23	yes	0–25	51.3	2.16	0.11	yes
chrF2	yes	0–100	100.0	9.78	8.96	no
chrF++	yes	0–100	100.0	<i>n/a</i>	13.91	no
CometKiwi-22	no	0–1	42.3	0.35	-0.10	yes
MetricX-24-QE	no	0–25	38.5	<i>n/a</i>	-0.42	yes

Table 2: **What each metric charges for an obligatory-slot error.** *ref.*: does the metric see a reference? *no* is reference-free (quality estimation). *win*: share of the panel’s own slot errors on which the metric ranks the correct form first (%); below 50 it prefers the wrong-valued output. $\times$ *agr.*: de-confounded slot signal as a signed multiple of the 75%-agreement threshold (Kocmi et al., 2024), negative meaning the metric prefers the wrong value; *n/a* where no threshold has been published for the metric. *scale*: the metric’s native score range. $\Delta_{av}\%$: mean penalty for the epicene form the systems emit, over 1252 items, as a percentage of that native range so that rows on different scales can be read against each other; the last column marks whether the 95% bootstrap CI on the raw penalty includes zero. **Read the last two columns by reference type.** The 2 reference-free metrics are below the rule: both win on under half the panel’s own errors and both have an avoidance interval containing zero, which is the paper’s claim. Most reference-based metrics above the rule *do* charge for avoidance, and their intervals exclude zero.

5 What Metrics Charge for an Obligatory-Slot Error

The scoring set, and the three arms. The metric study is scored on in-sentence-cue items only, over 6231 records drawn from the ten-system panel run on the superseded release (§8). We group them into three kinds of evidence, which we call *arms*. The *controlled* arm is constructed: a minimal pair differing in exactly one grapheme-cluster span, with the analyser confirming that the rejected member carries the wrong value. The *natural* arm is not constructed: the pair is an error the panel actually made and its repair, with nothing spliced. The *avoidance* arm is neither. Its emitted member is the epicene -ஆர் -*ār* form, which is not an error in the usual sense. **The three do not agree, so every per-slot number below names the arm it comes from.**

One benchmark limitation shapes the design. The released gold field is the *slot-bearing core*, a median of two words against five- to nine-word sources, so **the benchmark ships no full-sentence**

reference. We therefore use the DEMETR/ACES construction (Karpinska et al., 2022; Moghe et al., 2025). Because the good variant is then the reference itself, we also re-score against another system’s independent output, which takes COMET-22 from 100.0% to 92.1%. **Direction and magnitude are reported separately.** Magnitude is calibrated against Kocmi et al. (2024): the score difference at which human raters agree three times in four, which we call the **75%-agreement threshold**. 3 metrics have no published calibration and are left uncalibrated rather than given another checkpoint’s thresholds. Direction is Holm–Bonferroni corrected within family, one family per (cell $\times$ comparison), 257 families in total.

Our prediction was falsified for the neural reference-based metrics. We pre-registered that neural metrics would not charge a human-noticeable penalty for a single obligatory morpheme. That prediction is false for 3 of the 4 neural metrics that have a published calibration (Table 2): against that threshold, COMET-22 charges $8.52\times$, xCOMET-XL $6.55\times$ and MetricX-23 $2.16\times$; it holds for CometKiwi-22, and 2 neural metrics have no calibration key. chrF2 is a character *n*-gram metric, not a neural one, and the prediction never covered it; it charges $9.78\times$ as well, so the reference-based penalty is not a property of learned metrics. The signal we test it on is de-confounded: each metric also prefers unspliced text over spliced, so we difference the controlled and natural arms over the items they share, which removes that preference and leaves the slot signal. **“Neural metrics are blind to morpheme errors” is not true as a blanket claim.** The raw controlled-arm ratios before de-confounding are larger and are not the ones we defend.

On real errors, both reference-free metrics prefer the wrong output. This needs no splicing and no control slot. The *natural* arm is the 1479 errors the panel itself makes, each paired with its repair. Both members are well-formed Tamil; *wrong* means only that the slot carries the wrong value. On these pairs CometKiwi-22 ranks the correct form first 42.3% of the time and MetricX-24-QE 38.5%, so each *prefers the wrong output more often than not*. Every reference-based metric stays above chance on the same pairs; the closest is MetricX-23-large at 51.3%. Two checkpoints from different labs failing the same way makes this a property of the reference-free setting rather than of one model.

Avoidance is free, under reference-free evaluation. The panel’s dominant behaviour in the gender slot is not a wrong choice but the epicene -அர் *-ār*, a legitimate Tamil form that silently drops information the grammar requires. There are 1252 such outputs, all real panel outputs rather than constructed pairs. Over them CometKiwi’s mean penalty is -0.000965, with a 95% bootstrap interval from -0.00426 to 0.00275 that **includes zero**; that is 0.23 of the agreement threshold. MetricX-24-QE’s interval includes zero as well. **Under reference-free evaluation, a system that never marks the distinction loses nothing.** The qualifier is load-bearing and Table 2 carries it: the reference-based metrics do charge for avoidance, and their intervals exclude zero. The title is a claim about what reference-free evaluation can see, which is the setting most system-level monitoring runs in. Within that setting the result holds in both arms and does not depend on the control slot.

What a metric can rank tracks recoverability, and it is not exclusive to reference-free metrics. Figure 1 orders the slots by what an English source makes recoverable rather than by what it grammatically marks, because that is what the data order on. In the *controlled* arm CometKiwi-22 scores 99.2% on number, which English marks grammatically, against 80.2% on rationality and 79.1% on gender, which English leaves to the noun and the name. MetricX-24-QE gives 99.7% against 88.9% and 81.9%. **Three qualifications, and they are not small.** First, the ordering is a *controlled*-arm result and is gone in the natural arm, where CometKiwi scores 44.3% on number against 61.6% on rationality. Second, it is not exclusive to reference-free metrics: the reference-based MetricX-23-large separates the slots further (89.8% against 57.9%), while the equally reference-based xCOMET-XL is nearly flat (96.5% against 93.9%). Third, the two slots for which English has no grammatical exponent at all were the floor of this curve (CometKiwi 67.9% and 20.7%). Both are withdrawn for construction defects, and **their loss removes the strongest part of it.**

xCOMET detects the region and does not localise it. xCOMET advertises identification of “localized critical errors” (Guerreiro et al., 2024), which makes it the sharp target. Against the gold morpheme span, with numeral and named-entity perturbations of the *same* sentences as positive controls, the verdict is detects but does not lo-

calise. It flags the morpheme’s region on 91.1% of items, but that region covers 71.2% of the sentence against a gold span of 3.3 characters, so the hit rate at the conventional half-overlap threshold is 9.4%. It also flags 72.9% of *unperturbed* correct sentences. **The finding is severity.** When it flags the obligatory morpheme it calls it critical 6.3% of the time; on a named-entity swap in the same sentences, 67.7%; on a *duplicated word*, DEMETR’s canonical *minor* error, 52.4%. The control is what makes this specific: a metric merely weak on Tamil would not rate a duplicated word critical far more often than a violation that changes who the sentence is about, so DEMETR’s minor-versus-critical finding reproduces on Tamil. The obvious remedy, augmenting a metric’s training data, is bounded by Amrhein and Sennrich (2022) and Singh et al. (2024) (§9).

6 Model Behaviour, Provisionally

A prior-sentence cue suppresses avoidance without supplying the answer. If a reference-free metric cannot see the distinction, the next question is what the systems do with it. One result here is solid, and §3’s taxonomy is what makes it visible: because AVOIDANT is a first-class outcome rather than an error, commitment and accuracy move independently. In rationality, moving the cue out of the sentence and into the sentence before makes systems *more* willing to commit, not less: commitment rises on 12 of 13 systems, from 73.7–90.3% with an in-sentence cue to 66.2–100.0% with a prior-sentence one. Accuracy on the items they do commit to moves the other way, from 88.4–100.0% down to 49.6–95.1%. It holds across four prompted families and both dedicated MT systems, so no architecture story explains it. The context does not make these systems right; it stops them avoiding and makes them commit anyway (per-system rates: Table 6, Figure 2, Appendix B).

What does not hold, stated so it is not read into the table. That is the only cue-position claim the data support. Every model number here is measured on the corrected release, on 13 systems, and by the repaired checker: the -இன் past-class repair (Appendix C) landed after this panel was first scored, so the panel was re-scored with it; every rationality number here is post-repair, and the re-score moved no cell of any other slot. **The number rows carry no cue-position reading:** systems of-

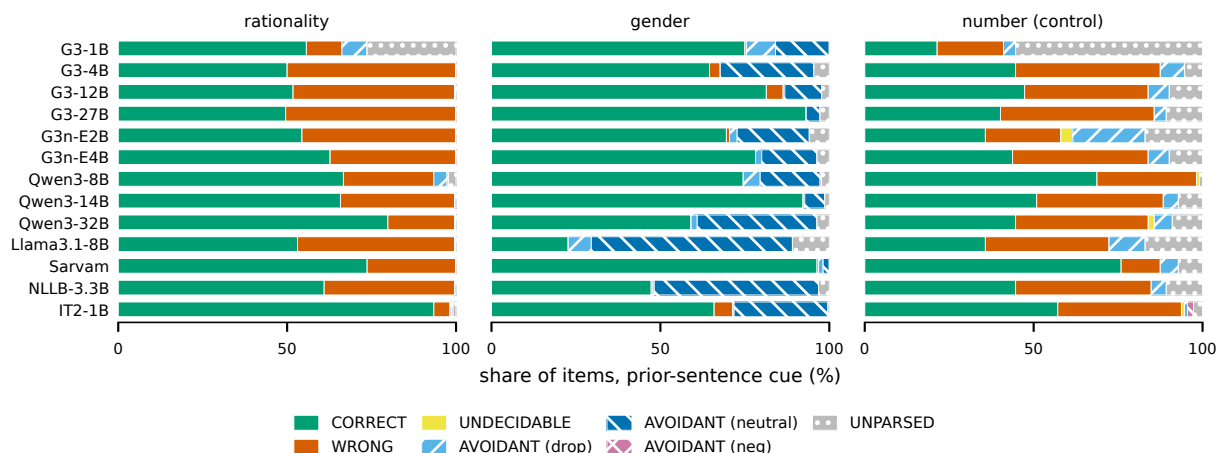


Figure 2: **Outcome composition with a prior-sentence cue.** Each bar is one system’s outcomes on one slot, as shares of its items: correct, wrong, avoidant, undecidable, unparsed. Avoidance is hatched and in its own hue family because it is a different event from an error: a system that declines to specify and a system that specifies wrongly are doing different things, and folding the two together would report an accuracy collapse where there is a refusal. Labels are ISO 15919.

ten answer with -ஆர் $\bar{a}r$, which is either honorific singular or plural, so the same output is scored wrong on one item and right on another and the instrument can move the difference by itself; **the number control does not function on these items** either, with accuracy among committed items under a prior-sentence cue between 46.9% and 86.7% against chance 50%, so **no between-slot comparison is drawn anywhere in this paper**; and **gender shows no reliable direction** between the two cue positions. One caveat bounds even the result above: the prior-sentence condition runs a single context prompt, so template interaction cannot be separated from cue position.

Deleting the cue is a different manipulation. Commitment falls from the in-sentence condition to the cue-free one on 13 of 13 systems in gender and on 0 in rationality (Table 9, Figure 5). What fills the gap is not the corpus: every system emits the same rationality value on nearly every item (100.0% for Gemma 3 27B) against a prior range of 34.5–69.6% (Table 7, Figure 4), so the defaults are real and system-specific but unexplained. Finally, the capability ceiling does not clear uniformly, and **the instruction that measured it was itself defective**: it named the benchmark’s internal label codes rather than the Tamil form the system was meant to produce, moving accuracy by up to 75.0 points (Table 8, both verdicts reported). Even under the corrected wording only 6 of 11 systems clear chance by a margin on gender and 0 of 11 on rationality, so a low rationality score cannot be distinguished

from incapacity.

7 Where the Decision Is Fixed

If a reference-free metric cannot see a distinction the source does not encode, the next question is whether the *model* settles it any earlier. We answer it causally on gemma-3-4b-pt, against a preregistration frozen before the first result file (checked by file modification time rather than asserted), by symmetric activation patching of the residual stream at the decision token in both directions (Zhang and Nanda, 2024). The statistic is the *onset*: the normalised depth at which patching first restores 0.50 of the clean logit difference and stays above it, bootstrapped over templates. **Number is the control** because Ferrando and Costa-jussà (2024) report a language-independent subject-number direction in a Gemma base model, so a null there would mean a broken instrument.

The marked distinction is fixed halfway; the unmarked one is not. Table 3 gives the onsets. Number, which English marks, reaches onset at 0.561 of normalised depth; rationality, which English leaves unmarked, not until 0.888. The gap that matters is rationality minus number: 0.327, with a 95% CI of 0.301–0.339. It excludes zero and clears the pre-registered floor of 0.200, and both patching directions agree. **The control reproduces**: peak restored effect on number is 0.992, which is Ferrando and Costa-jussà’s positive result on our items and makes this a *dissociation* rather than a bare null. In this checkpoint, a distinction

Slot	marked in English	onset (95% CI)
number (<i>control</i>)	yes	0.561 (0.550–0.585)
gender	no	0.731 (0.710–0.863)
rationality	no	0.888 (0.878–0.891)
difference		0.327 (0.301–0.339)

Table 3: **Where each decision is causally fixed in gemma-3-4b-pt.** Onset is the normalised depth at which activation patching first restores 0.50 of the clean logit difference and stays above it, bootstrapped over templates. The last row is rationality minus number, the pre-registered contrast; its interval excludes zero and clears the pre-registered floor of 0.200. Number is the control: Ferrando and Costa-jussà (2024) report a subject-number direction in a Gemma base model, and it reproduces here.

English marks is causally settled around the middle of the network; one English leaves unmarked is not settled until close to the output.

The probe study failed its own pre-registered bar, and we report that. The same frozen document specified a layerwise linear probe that had to add 0.75 AUC over surface-token controls. It does not: the peak excess is 0.29 for rationality, 0.18 for number and 0.02 for gender, so the kill criterion fires, and we drop the probe rather than override it. That same document made patching the primary instrument and the probe the secondary one, which is why the headline result survives the failure: patching locates where the decision is causally fixed, a probe only where a classifier finds a correlate. What the failure costs is real. We make no decodability claim, and the tuned lens that would separate “not represented” from “represented and not yet decoded” was not run, so that alternative is untested rather than excluded.

8 Three Confounds in One Slot

All of the above is measured on the corrected release; this section is how it came to need correcting. The first release produced a clean, quotable result on the gender slot: the language models almost never committed to a value, while the dedicated MT systems did. **We retract it.** It was a property of how the items were built, and the three defects behind it are independent.

The claim does not weaken, it inverts. On the corrected items the ordering is the other way round. The 11 prompted systems commit at a median of 75.4% on prior-sentence gender against 47.7–

System	superseded	corrected	
	commit	commit	acc.
Gemma 3 1B	51.8	75.4	99.4
Gemma 3 4B	2.3	67.7	95.3
Gemma 3 12B	0.0	86.4	94.2
Gemma 3 27B	0.0	93.2	100.0
Gemma 3n E2B	<i>n/a</i>	70.5	98.7
Gemma 3n E4B	<i>n/a</i>	78.2	100.0
Qwen3 8B	2.3	74.6	100.0
Qwen3 14B	0.4	92.3	100.0
Qwen3 32B	<i>n/a</i>	59.1	100.0
Llama 3.1 8B	0.0	22.7	<i>n/a</i>
Sarvam-Translate	0.0	96.8	99.5
NLLB-200 3.3B	22.7	47.7	99.1
IndicTrans2 1B	44.1	71.4	92.4

Table 4: **The gender result was a property of the items, not of the systems.** Prior-sentence gender commitment (%) on the superseded release, whose prior-sentence source is singular *they*, and on the corrected one, whose prior-sentence source is a singular animate name. The same systems move by up to 96.8 points; every other slot moves by at most 26.8. A commitment cell prints *n/a* where that system was not run on the superseded release; the accuracy cell prints *n/a* below the reporting floor (< 30 committed items or < 30% commitment). **The original claim does not merely weaken, it inverts:** the 11 prompted systems reach a median of 75.4% against 47.7–71.4% for the 2 dedicated MT systems, and 7 of the 11 exceed every one of them (§8).

71.4% for the 2 dedicated MT systems, and 7 of the 11 exceed every one of them: Sarvam-Translate 96.8% and Gemma 3 27B 93.2% against NLLB-200 47.7% and IndicTrans2 71.4% (Table 4 is the per-system difference). We state the inversion rather than softening the claim: its sign is wrong, not merely its support.

The three defects, each independent of the others. A neutral anaphor: singular *they* against a singular-feminine gold, in the one slot whose Tamil target has a neutral exponent, so a *correct* translation scored as avoidance. Two cued conditions that differ in cue *type* rather than distance, the prior-sentence one being the *easier* on 11 of 13 systems: **no comparison between them may be read off this benchmark, in either release**, which is why §6 compares in-sentence against cue-free instead. And a leak detector keyed on output length, blind to a system that translates the context *instead of* the marked sentence: the dangerous shape, since it is in the context that a prior-sentence cue lives. Appendix F gives each in full.

The audit, and why the defective release is kept.

All three were found by an automated check that asks, per slot $\times$ condition, whether the source carries *any* exponent of the feature the gold commits to — deliberately lexical rather than clever, since the weakest question that would have caught them is the one worth running on every release. It returns 1141 unanswerable items in the first release across 12 (slot, condition, check) categories and 0 in the second, where it is a hard check. **The defective release is retained as a positive control**, so every guard is verified against a real defect.

9 Related Work

The nearest neighbour. GENDER1PERSON (Popović and Lapshinova-Koltunski, 2025) shares the experimental logic exactly: English leaves first-person gender unmarked, the target obligates it, and the skew is measured on system output. Three things separate this paper from it. Scope: gender in two Indo-European targets there, a Dravidian agreement system here, where one slot (rationality) has no prior MT treatment we could find. The informed conditions: both cued conditions carry gold, so a skew that survives when the answer is available is a prior overriding evidence rather than a gap being filled. And, the contribution we would defend first, it does not ask what *evaluation* charges for the error, which is §5’s question. GRAPE (Menis Mastromichalakis et al., 2025), Staliunaite et al. (2026), Dawkins et al. (2025) and Prates et al. (2020) develop the same programme with external social statistics, entropy, ambiguity preservation and output proportions.

DEMETER (Karpinska et al., 2022) licenses the question and explains why it cannot answer it here: all its examples are into English. ACES (Moghe et al., 2025) is the largest challenge set for MT metrics and its taxonomy is the MQM *accuracy* branch, so morphology is absent by construction and English→Tamil is 1 of its 36476 rows; IndicMT Eval (Sai B et al., 2023) supplies the Tamil half, where metrics fail on the fluency branch that obligatory-slot errors land on. Amrhein and Sennrich (2022) and Singh et al. (2024) bound how far retraining closes a metric’s blind spot, and both targeted a phenomenon that *is* marked in the source, so supervision could be derived automatically: which an unmarked distinction does not allow. Kocmi et al. (2024) supplies the calibration, Guerreiro et al. (2024) is the localisation claim §5

tests, and Robinson et al. (2023) make resource level the dominant predictor of translation *quality*, an axis we argue is orthogonal to obligatory-slot correctness.

The paradigm runs from LingEval97 (Sennrich, 2017) through Bawden et al. (2018), whose balance constraint our prior-sentence condition adopts directly, and ContraPro (Müller et al., 2018), to the DFKI WMT suites (Avramidis et al., 2020). Morpheval (Burlot and Yvon, 2017; Burlot et al., 2018) is the sharpest contrast: it perturbs a distinction that *is* marked in the source and checks whether it propagates, though its hyponym-substitution family is the nearest precedent for our cue-free condition. Contrastive conditioning (Vamvas and Sennrich, 2021a) detects which reading a system defaults to under *source ambiguity*; our slots have none. CoCoA-MT (Nadejde et al., 2022) and the IWSLT formality tasks (Anastasopoulos et al., 2022; Agarwal et al., 2023) run “unspecified” as an uncontrolled baseline row rather than a condition, and Měchura (2022) named the category without measuring any system’s defaults. The reference grammars, treebanks and analyser papers behind our Tamil facts are listed in appendix A.

10 Conclusion

Our pre-registered prediction that neural metrics cannot see a single obligatory morpheme is false for the neural reference-based metrics, and we report the falsification. What survives is narrower and harder to explain away: on the errors these systems actually make, both reference-free metrics rank the wrong output first more often than not, and under both, the epicene form that drops the distinction carries a penalty whose interval includes zero. Reference-based metrics do charge for it, so the claim is about what reference-free evaluation can see.

Limitations

This is one language in one register: written Tamil, and a Dravidian agreement system whose particular shape may not generalise to other target languages that obligate distinctions English leaves unmarked. The panel is open-weight models of at most 32B parameters, so we cannot say whether a frontier system behaves differently. Three measurement limits bound what we claim rather than qualifying it. First, the number slot is withheld

from every cue-position result: systems often answer with -ஆர் *-ār*, which is either honorific singular or plural, so the same output string is scored wrong on a singular item and right on a plural one, and any effect there would be measuring the instrument. Second, one prompt supplies the prior-sentence condition, so template interaction cannot be separated from cue position; the direction is consistent across systems, but the magnitude is not attributable to cue position alone. Third, the rationality analyser was repaired after the panel was first scored and the panel was re-scored with the repair; the repair alone moved in-sentence rationality commitment by up to 21.1 points on one system (Appendix C), which is a measure of how far one analyser defect can move a headline cell, and a defect not yet found could move it again. The pre-repair scoring is released with the repaired one. The probe study reports a pre-registered failure alongside its causal result: the layerwise probe never cleared its frozen threshold, so we establish *where* a distinction is causally fixed without establishing that it is linearly decodable earlier, and the tuned lens that would separate “not represented” from “represented but not decoded” was not run. The benchmark was verified by two native speakers of Tamil, whose rulings corrected shipped data in two respects (second-person address is binary, and the neuter singular is ungrammatical on a plural subject). Verification is described throughout as human-led with finite-state support, never as verified by the transducer alone: round-tripping a form through the analyser that generated it establishes internal consistency, not correctness, and two forms that round-tripped cleanly proved to have no corpus attestation at all.

Ethics and Broader Impact

Choosing the wrong value in an obligatory slot is not a style error. Assigning அஃறிணை *akṛiṇai* (the agreement class Tamil uses for animals, plants and objects) to a human referent is a dehumanising error in a way a fluency slip is not, and the finite verb forces the choice in every clause. The harm is asymmetric: it lands on the reader in the target community, who cannot see the source, cannot tell that the source was silent, and has no way to attribute the offence to the system rather than the speaker. A translation that is fluent, adequate and highly scored by every automatic metric can be socially unusable for its entire length. The gender

slot is a site of misgendering harm and is present for commensurability with the gender-bias literature rather than as a claimed contribution; we retract the behavioural claim our first release made about it (§8). A system emitting the epicene form commits no misgendering and also destroys information the source did supply, and whether that is preferable is a judgment about deployment context rather than a fact about the model, which is why avoidance is a separate outcome here and not an error, and why we publish gold only where the source determines the answer.

Annotators, data and provenance. The two native speakers who verified the benchmark were informed of the use of their judgments and are credited or anonymised at their choice. Released data are CC BY-SA 4.0, forced by the ShareAlike term on UD_Tamil-MWTT rather than chosen; code is Apache-2.0. UD_Tamil-TTB is CC BY-NC-SA 3.0, so it is excluded from the release path and only aggregate counts derived from it appear here.

References

- Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cetolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declercq, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, and 43 others. 2023. [FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Chantal Amrhein and Rico Sennrich. 2022. [Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.
- Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marceley Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, and 24 others. 2022. [Findings of the IWSLT 2022 evaluation campaign](#). In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin,

- Ireland (in-person and online). Association for Computational Linguistics.
- E. Annamalai and Sanford B. Steever. 1998. Modern Tamil. In Sanford B. Steever, editor, *The Dravidian Languages*, pages 100–128. Routledge, London, UK.
- A. H. Arden. 1942. *A Progressive Grammar of Common Tamil*, fifth edition. Christian Literature Society, Madras, India.
- Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. [Fine-grained linguistic evaluation for state-of-the-art machine translation](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.
- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. [Evaluating discourse phenomena in neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.
- Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. [The WMT’18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.
- Franck Burlot and François Yvon. 2017. [Evaluating the morphological competence of machine translation systems](#). In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.
- Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2025. [Gender-neutral machine translation strategies in practice](#). In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 74–88, Geneva, Switzerland. European Association for Machine Translation.
- Bonnie J. Dorr. 1994. [Machine translation divergences: A formal description and proposed solution](#). *Computational Linguistics*, 20(4):597–633.
- Javier Ferrando and Marta R. Costa-jussà. 2024. [On the similarity of circuits across languages: a case study on the subject-verb agreement task](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. [DEMETER: Diagnosing evaluation metrics for translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. [Navigating the metrics maze: Reconciling score magnitudes and accuracies](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.
- Parameswari Krishnamurthy and Kengatharaiyer Sarveswaran. 2021a. [Towards building a modern written Tamil treebank](#). In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 61–68, Sofia, Bulgaria. Association for Computational Linguistics.
- Parameswari Krishnamurthy and Kengatharaiyer Sarveswaran. 2021b. [UD_Tamil-MWTT: Universal Dependencies Tamil Modern Written Tamil Treebank](#). Universal Dependencies treebank release.
- Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge University Press, Cambridge, UK.
- Thomas Lehmann. 1993. *A Grammar of Modern Tamil*, second edition. Pondicherry Institute of Linguistics and Culture, Pondicherry, India.
- Michal Měchura. 2022. [A taxonomy of bias-causing ambiguities in machine translation](#). In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington. Association for Computational Linguistics.
- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. [Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational](#)

- terms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32233–32249, Suzhou, China. Association for Computational Linguistics.
- Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. [Machine translation meta evaluation through translation accuracy challenge sets](#). *Computational Linguistics*, 51(1):73–137.
- Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. [A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.
- Dineskumar Murugesapillai, Anankan Ravinthirasa, Gihan Dias, and Kengatharaiyer Sarveswaran. 2021. [Neural-based Tamil grammar error detection](#). In *Proceedings of the First Workshop on Parsing and its Applications for Indian Languages*, pages 27–32, NIT Silchar, India. NLP Association of India (NLP AI).
- Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. [CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.
- P. Perumalsamy. 1991. *Status Marking in Tamil: A Sociolinguistic Study*. Ph.D. thesis, Bharathiar University, Coimbatore, India. Published in *Language in India* 8(6), 2008, <https://www.languageinindia.com/june2008/profperumalsamypdhddissertation.pdf>.
- Maja Popović and Ekaterina Lapshinova-Koltunski. 2025. [GENDER1PERSON: Test suite for estimating gender bias of first-person singular forms](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 800–822, Suzhou, China. Association for Computational Linguistics.
- Kabilan Prasanna and Aryaman Arora. 2024. [IruMozhi: Automatically classifying diglossia in Tamil](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3096–3103, Mexico City, Mexico. Association for Computational Linguistics.
- Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. 2020. [Assessing gender bias in machine translation: a case study with Google Translate](#). *Neural Computing and Applications*, 32(10):6363–6381.
- Vinodh Rajan. 2018. [Aksharamukha: Script converter and transliterator](#). Software.
- Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. [Prague dependency style treebank for Tamil](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 1888–1894, Istanbul, Turkey. European Language Resources Association (ELRA).
- Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.
- Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. [IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.
- Kengatharaiyer Sarveswaran. 2024. [Morphology and syntax of the Tamil language](#). *Preprint*, arXiv:2401.08367. CC BY-SA.
- Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. [ThamizhiMorph: A morphological parser for the Tamil language](#). *Machine Translation*, 35(1):37–70.
- Harold F. Schiffman. 1999. *A Reference Grammar of Spoken Tamil*. Cambridge University Press, Cambridge, UK. The 1996 web precursor, *A Grammar of Spoken Tamil*, is at <http://ccat.sas.upenn.edu/plc/tamilweb/book.html>.
- Rico Sennrich. 2017. [How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.
- Anushka Singh, Ananya Sai, Raj Dabre, Ratish Pudupully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. [How good is zero-shot MT evaluation for low resource Indian languages?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.
- Ieva Staliunaite, Julius Cheng, and Andreas Vlachos. 2026. [Uncertainty quantification for evaluating gender bias in machine translation](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2204–2225, Rabat, Morocco. Association for Computational Linguistics.
- S. Suseendrarajah. 1993. *Jaffna Tamil: Phonology and Morphology*. University of Jaffna, Jaffna, Sri Lanka.

Jannis Vamvas and Rico Sennrich. 2021a. [Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021b. [On the limits of minimal pairs in contrastive evaluation](#). In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.

A Reproducibility

We release the template generator, the 2873-item benchmark with its gold targets, the FST-based checker with the five analyser repairs of Appendix C, the prompt set, and the raw outputs and scored records of all 13 systems, under the licences given in the ethics statement. Both the pre-repair and the repaired scoring are included, so the re-score reported in Appendix C can be checked record by record. Every run records its resolved model revision, decoding configuration and seed (20260807), so any panel system can be re-run and re-scored against the same gold; the withdrawn slots and the superseded first release of the set are retained alongside the corrected one rather than deleted (§8).

Scoring policy. Determinism is why the checker is an FST rather than a neural model: the best published neural Tamil grammatical-error system reports F1 73.4 (Murugesapillai et al., 2021). Scoring is universal-after-context-filtering: only analyser readings compatible with the item’s declared referent number are retained, and the form counts as correct only if *all* surviving readings match the target. The existential policy is computed as a robustness column and the residual undecidable rate carries its own denominator. Of the 3980 items built, 3040 carry gold; 4 are marked undecidable and are emitted with the marker rather than dropped. The released gold field is the column `gold_targets`. Direction tests in §5 are Holm–Bonferroni corrected within family across 257 families; magnitude intervals are cluster bootstraps over `template_id`, so items sharing a template are not counted as independent.

Sources for the Tamil facts. The grammatical descriptions this paper relies on come from Lehmann (1993); Schiffman (1999); Arden (1942); Perumalsamy (1991); Annamalai and Steever (1998); Suseendrarajah (1993); Krishnamurti (2003); Prasanna and Arora (2024); Dorr (1994), the annotation facts of §1 from Ramasamy and Žabokrtský (2012); Krishnamurthy and Sarveswaran (2021a,b); Batsuren et al. (2022), the analyser from Sarveswaran et al. (2021); Sarveswaran (2024); Murugesapillai et al. (2021) and the positive control §7 reproduces from Ferrando and Costa-jussà (2024).

Figures. Every series carries colour, line style and marker redundantly, so the figures survive greyscale printing. Figure labels are ISO 15919 romanisation because the plotting library has no complex-text shaper for Tamil; Tamil script appears in tables and body prose only.

One methods result worth reusing. We first cached the residual stream in `float16`. **On Gemma-3 that overflows:** the residual passes the `float16` maximum of 65504 in the upper third of the network: exactly the depths this result is about. On 200 real Tamil chat-formatted prompts `float32` matches the reference implementation on 200 top-1 predictions at a maximum logit-difference error of 4.67e-05, while `bfloat16` matches 197 at 0.25 larger than several of the effects above. The study runs `float32` throughout.

The two withdrawn slots. §2 gives the honorificity mechanism. The imperative sub-family’s operative targets also carry no exponent the analyser will license. Clusivity has no valid in-sentence condition at all: only 0 of 196 unique in-sentence sources contain a first-person-plural English pronoun.

B Full Per-System Results

Table 6 is the per-system commitment rate §6 reports and Figure 3 the same rates against model scale; the outcome composition behind them is Figure 2 in §6, and the per-system superseded-against-corrected difference on which the retraction rests is Table 4 in §8.

Where the systems fall back to. Table 7 and Figure 4 give the emitted distribution on cue-free items against the between-corpus prior, per system and per slot.

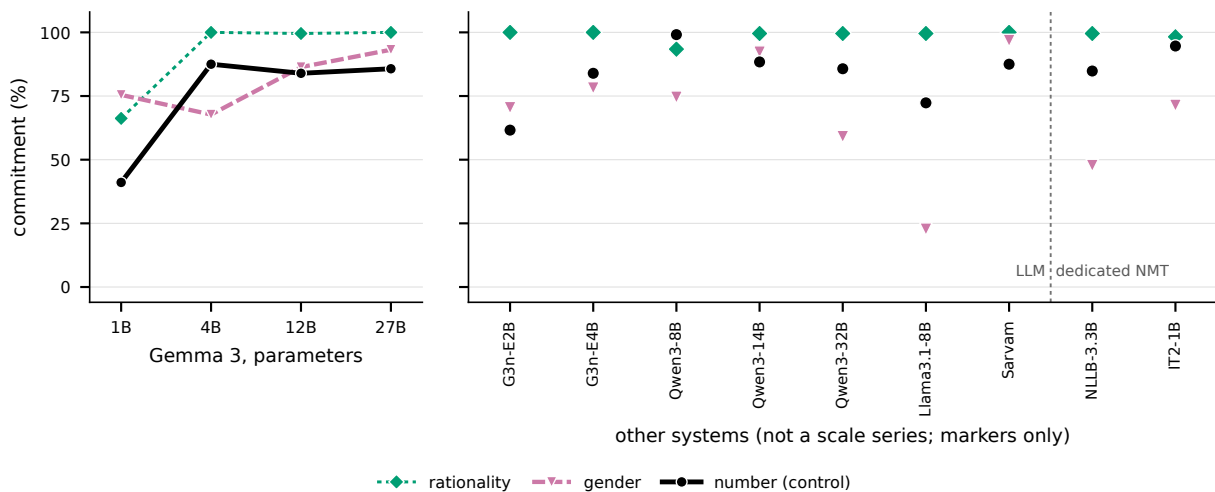


Figure 3: **Commitment rate by slot with a prior-sentence cue, on the corrected release.** Commitment is the share of items on which a system emits a form that takes a value in the slot. The four Gemma 3 sizes on the left are a scale series and are joined; the right-hand systems are markers only, are not a scale series, and are not joined by a line. Labels are ISO 15919; matplotlib cannot shape Tamil.

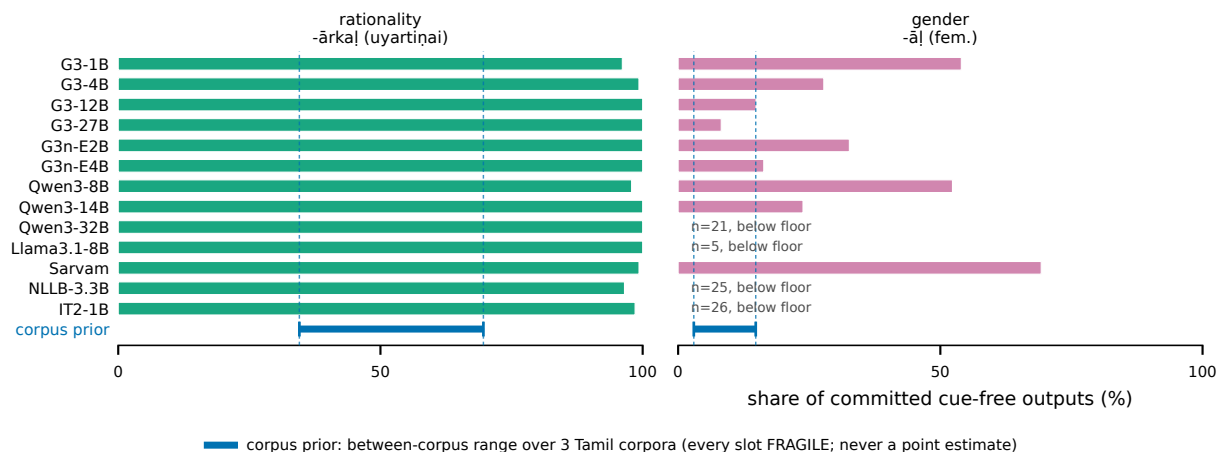


Figure 4: **Cue-free output against the corpus prior, drawn as a range.** Every slot is fragile by a criterion fixed in advance — which corpus is chosen moves the prior more than how many tokens are counted, so the prior is a band and never a line. In rationality every system’s rate falls outside the band, some above its top and some below its bottom.

The capability ceiling. The instructed condition names the target value in the instruction, so a system failing it cannot be read as defaulting in any of the other three. The shipped instruction was defective — it named the benchmark’s internal label codes rather than the Tamil morpheme, and naming the actual target form moves accuracy by up to 75.0 points. Both verdicts are reported side by side across all 55 cells rather than one replacing the other (Table 8). Even under the corrected wording the ceiling is not cleared everywhere, and §6 states which slots it fails for.

Deleting the cue. Table 9 and Figure 5 are the in-sentence-against-cue-free comparison §6 reports,

per system and per slot.

C Analyser Repairs

Off-the-shelf Tamil morphological tooling required four documented repairs before it could score obligatory-slot agreement, and a fifth landed after the panel reported here was first scored; the panel was re-scored with it, and the paragraph after the list reports what that moved. Each is at the scoring layer; the validated transducer is untouched, and both ablations are published (Table 10).

1. **-ஆர் -ār syncretism.** The suffix is honorific-singular or rational-plural and the transducer is inconsistent about which. Resolved by the

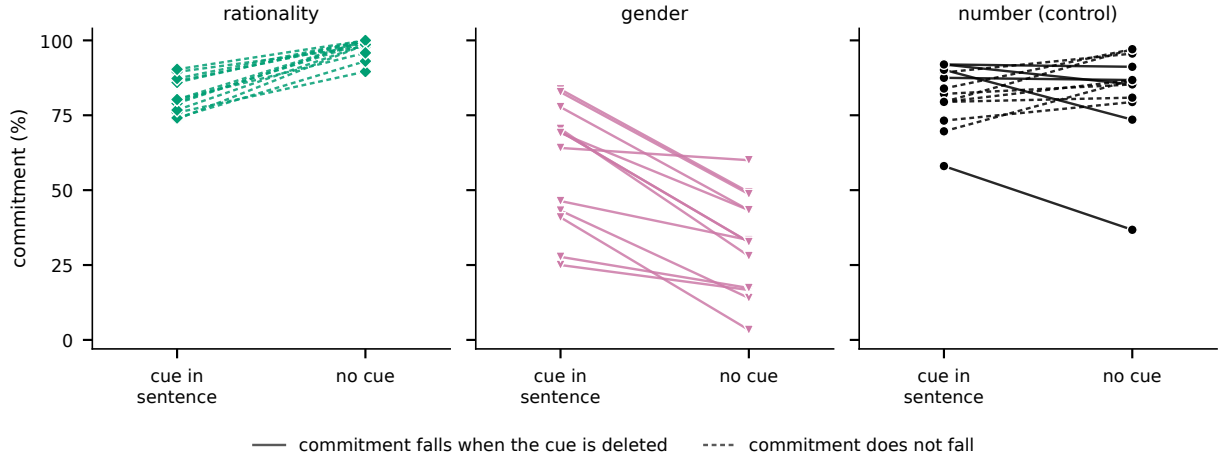


Figure 5: **From an in-sentence cue to no cue: commitment when the cue is deleted.** Each line is one system; commitment is the share of items on which it emits a form that takes a value in the slot. The cue-free condition carries no gold, so the quantity is commitment, not accuracy. The two cued conditions are not compared: they cue the value through different channels and the prior-sentence one is the easier on 11 of 13 systems (§8). Gender is the only slot in which deleting the cue reduces commitment across the panel: it falls on 13 of 13 systems there, against 0 in rationality and 5 in number.

Slot and Tamil realisation	Prior MT
Rationality (திணை) வந்தார்கள் / வந்தன <i>vantārkaḷ / vantaṇa</i>	none <i>prov.</i>
Gender in the verb வந்தான் / வந்தாள் / வந்தார் <i>vantāṇ / vantāḷ / vantār</i>	heavily covered <i>calib.</i>
Number (<i>control</i>) வந்தது / வந்தன <i>vantatu / vantaṇa</i>	marked in both <i>control</i>

Table 5: The scored slots, with the ISO 15919 scoring key under each Tamil form: the agreement features Tamil realises on the finite verb. Tamil is typeset by Xe-TeX with the Tamil shaper engaged; the romanisation is what appears in every figure, because `matplotlib` runs no complex-text shaper. **Rationality has no prior MT coverage at all** and is marked *provisional* (see Limitations); gender is a *calibration* axis, not a claimed contribution, and is the slot whose first measurement we retract (§8); number is the control, marked in both languages.

context filter, with the residual undecidable rate reported.

2. **-ஈர்கள் -īrkaḷ and -உங்கள் -uṅkaḷ.** Second-person plural suffixes are syncretic between plain-plural and honorific-singular, exactly as -ஆர்/-ஆர்கள் are in the third person, but are tagged bare. This is a grammatical claim we make and the transducer does not; it is in the native-check queue for that reason.
3. **-அனார் -aṇār.** Tagged singular although it is

unambiguously plural.

4. **Word-final sandhi.** A sandhi consonant hides an overt pronoun, so எங்களைப் *eṅkaḷaip* returns a spurious verb reading and inflates the AVOIDANT_DROP rate.
5. **-இன் -iṇ past, finite reading.** ஓடின *ōḍiṇa* is a finite verb and நாய்கள் ஓடின *nāykaḷ ōḍiṇa* is a complete sentence, but the transducer admits only +nonfin+adjpart, so the checker scored correct Tamil as wrong. The repair adds the finite reading and keeps the participial one, since the form is genuinely ambiguous. It landed after the reported panel was first scored.

What the fifth repair moved. The panel was re-scored from the same raw outputs with the repair on, every rationality number in this paper is from that re-score, and the pre-repair scoring is released beside it. Of 54443 scored records, 728 changed outcome, 725 of them in the rationality slot and 3 elsewhere; no headline cell of any other slot moved. 717 rationality outcomes became CORRECT (512 from UNPARSED, 139 from WRONG), every one of them on அஃறிணை gold, where the -இன plural is the target form; 8 UNPARSED outputs on உயர்திணை gold became WRONG. 568 of the changes are in the in-sentence condition, and all 13 systems are affected: in-sentence rationality commitment moved from 58.3–75.0% to 73.7–90.3%, a rise of up to 21.1 points on one system. The old analyser was *under-reporting* commitment, not

System	rat.	gen.	num.
Gemma 3 1B	66.2	75.4	41.1
Gemma 3 4B	100.0	67.7	87.5
Gemma 3 12B	99.6	86.4	83.9
Gemma 3 27B	100.0	93.2	85.7
Gemma 3n E2B	100.0	70.5	61.6
Gemma 3n E4B	100.0	78.2	83.9
Qwen3 8B	93.4	74.6	99.1
Qwen3 14B	99.6	92.3	88.4
Qwen3 32B	99.6	59.1	85.7
Llama 3.1 8B	99.6	22.7	72.3
Sarvam-Translate	100.0	96.8	87.5
NLLB-200 3.3B (NMT)	99.6	47.7	84.8
IndicTrans2 1B (NMT)	98.2	71.4	94.6

Table 6: **Commitment rate with a prior-sentence cue (%)**: does the system make the obligatory choice at all? A system *commits* when it emits a form that takes a value in the slot. Measured on the **corrected** release, on which every system in the original panel has been re-run; the superseded release appears only in Table 4, where differencing the two is the point. All cells are the same context prompt for prompted systems; dedicated MT systems take no prompt, and they are below the rule. Cells remain *provisional* for the reasons §6 gives — the number control does not function on these items, the capability ceiling is not cleared for rationality, and the checker has a known defect in the rationality slot — not because any system is unmeasured. **Commitment, not accuracy**: accuracy is undefined at zero commitment and is reported separately in Table 9.

over-reporting error. The metric study is unaffected: re-judged with the repaired analyser, all 73 natural-arm and all 1415 controlled-arm rationality errors remain errors, and every gold-bearing variant remains correct.

D Annotation Facts

Table 11 is the full version of the counts §1 opens with. Each row is a direct query over the released treebanks and can be re-run against a later treebank release, so a release that changed any of them would change the table rather than the prose.

E Open Native-Speaker Queue

The queue of open native-speaker questions is generated from the released data rather than maintained by hand, and its size is reported rather than summarised. 134 rows are open in total: 26 design questions, 45 verb lemmas, 35 noun lemmas and 36 personal names, the last of which no Tamil transducer covers at all. 156 items are scored against a reading the analyser does not license, all of them in the withdrawn imperative honorificity family.

System	emitted (%)	corpus prior range (%)	obs. / prior range
<i>rationality</i>			
Gemma 3 1B	96.1	34.5–69.6	0.38–1.78
Gemma 3 4B	99.3	34.5–69.6	0.43–1.87
Gemma 3 12B	100.0	34.5–69.6	0.44–1.9
Gemma 3 27B	100.0	34.5–69.6	0.44–1.9
Gemma 3n E2B	100.0	34.5–69.6	0.44–1.9
Gemma 3n E4B	100.0	34.5–69.6	0.44–1.9
Qwen3 8B	97.9	34.5–69.6	0.41–1.83
Qwen3 14B	100.0	34.5–69.6	0.44–1.9
Qwen3 32B	100.0	34.5–69.6	0.44–1.9
Llama 3.1 8B	100.0	34.5–69.6	0.44–1.9
Sarvam-Translate	99.3	34.5–69.6	0.43–1.88
NLLB-200 3.3B	96.5	34.5–69.6	0.39–1.79
IndicTrans2 1B	98.6	34.5–69.6	0.42–1.85
<i>gender</i>			
Gemma 3 1B	54.0	3.0–14.8	2.65–17.07
Gemma 3 4B	72.2	7.0–22.3	2.24–9.28
Gemma 3 12B	85.1	7.0–22.3	2.82–11.12
Gemma 3 27B	91.8	7.0–22.3	3.12–12.06
Gemma 3n E2B	67.3	7.0–22.3	2.02–8.59
Gemma 3n E4B	83.7	7.0–22.3	2.75–10.91
Qwen3 8B	52.3	3.0–14.8	2.53–16.5
Qwen3 14B	76.2	7.0–22.3	2.42–9.85
Sarvam-Translate	69.2	3.0–14.8	3.68–22.16

A slot appears here only where a system’s cue-free cell clears the reporting floor *and* the slot has a reportable corpus prior; **number** has no prior of its own.

Table 7: **Emitted share on cue-free items**: the share of the modal value against the **between-corpus range** of the prior over 3 Tamil corpora. Every slot is fragile by the pre-registered criterion — which corpus one picks moves the prior more than how many tokens one counts — so the prior is a range and never a point estimate, and the observed-to-prior ratio is computed at both ends of that range. This licenses the statement that the output rate *deviates from general Tamil text frequency*; it does *not* license ‘deviates from its own training frequency’, because no evaluated system’s training mixture was measured. In rationality every system’s rate falls outside the corpus range, some above its top and some below its bottom.

Every lemma has its own row because round-tripping through the transducer is not evidence of correctness. A form the transducer over-generates round-trips perfectly, and at least one such form survived all four automatic screens and was still wrong: the -இன் *-in* past plural the transducer builds by concatenation, ஒடினன *ōṭinaṇa*, which has zero occurrences in 105 million tokens of Tamil while the real form ஒடின *ōṭina* is frequent. This is why the validation protocol is described as human-validated with FST support throughout, and why the surviving slots stay provisional. It also bounds the attestation gate in the other direction: the optatives இறங்குக *irāṅkuka*, ஏறுக

ēruka, கழுவுக *kaḷuvuka* and நீந்துக *nīntuka* are ordinary written Tamil with zero occurrences in the same corpora, so a zero count is evidence about our corpora and not about the language. The frequent optative வருக *varuka* is attested there and is not one of them.

F The Construction Audit

The audit behind §8 asks one question per (slot, condition): does the English source carry *any* exponent of the feature the Tamil gold commits to? It parses no English and inspects word inventories, which is deliberate — the weakest question that would have caught the defects is the one that can be run on every release and disputed word by word. It is advisory on the superseded release, which is retained as its positive control, and a hard check from the corrected release on: a change to the audit that stopped it detecting the original defect would be caught rather than pass quietly.

The three gender confounds in full. 1. **The anaphor was neutral in the one slot with a neutral target form.** 100.0% of prior-sentence gender sources used singular *they* against a singular-feminine gold, and gender is the one slot whose Tamil target admits the epicene -ஆர் *-ār*, so a neutral rendering was a *correct translation* that the scorer recorded as avoidance. On corrected items, whose prior-sentence source is a singular animate name, the same systems commit at 22.7–96.8% with 92.4–100.0% accuracy; nothing about the models changed. 2. **The two cued conditions differ in cue type, not cue distance.** The in-sentence condition cues gender through the onomastics of the Tamil name, which is world knowledge; the prior-sentence condition cues it through an explicit English appositive, which is reading comprehension. If the two differed only in where the cue sits, the in-sentence condition could not be the harder one. It is: the prior-sentence condition scores higher on 11 of 13 systems, and the per-system gap runs from -18.2 to 43.6 points. 3. **The leak detector was watching the wrong signal.** The shipped detector keyed on output length, so it was blind to a system that translates the context *instead of* the marked sentence — the dangerous shape, since it is in the context that a prior-sentence cue lives. Its replacement differences the rate at which the marked sentence’s own gold verb stem appears in the output, within a system so that paraphrase cancels against that system’s own base-

line. It exposes the failure at once: gemma3-1b misses that stem on 98.2% of prior-sentence gender items against 38.6% with an in-sentence cue, while every other system stays above 55.0% in both. On the superseded release, only 3 of the 5 slots the audit covers have an interpretable comparison between the two cued conditions; which do not is computed from the sources (clusivity, gender).

Three further defects. Besides the three gender confounds of §8, the audit found three defects nobody had reported: a condition whose English sources were built by deleting the cue string out of another condition’s frame, and which were therefore malformed; items whose Tamil reference named a person the English source never mentions, so the reference is not a translation of its source; and personal names used as gendered cues that appear in both gender sections of the name lexicon.

System	Slot	shipped (%)	corrected (%)
Gemma 3 12B	gender	70.0	95.0
	number	25.0	100.0
	rationality	50.0	50.0
Gemma 3 1B	gender	12.5	25.0
	number	0.0	25.0
	rationality	42.5	45.0
Gemma 3 27B	gender	80.0	100.0
	number	50.0	100.0
	rationality	50.0	55.0
Gemma 3 4B	gender	40.0	35.0
	number	0.0	25.0
	rationality	50.0	50.0
Gemma 3n E2B	gender	47.5	42.5
	number	50.0	75.0
	rationality	50.0	50.0
Gemma 3n E4B	gender	47.5	45.0
	number	50.0	50.0
	rationality	50.0	52.5
Llama 3.1 8B	gender	40.0	67.5
	number	25.0	50.0
	rationality	55.0	45.0
Qwen3 14B	gender	80.0	85.0
	number	25.0	75.0
	rationality	50.0	50.0
Qwen3 32B	gender	65.0	100.0
	number	50.0	50.0
	rationality	50.0	52.5
Qwen3 8B	gender	50.0	92.5
	number	35.0	100.0
	rationality	50.0	52.5
Sarvam-Translate	gender	47.5	37.5
	number	25.0	25.0
	rationality	50.0	50.0

Table 8: **The capability ceiling was measured with a defective instruction, and both verdicts are reported.** The shipped instruction named the benchmark’s internal label codes rather than the target Tamil morpheme. With an instruction that names the actual form, accuracy moves by up to 75.0 points. The instructed condition gates the interpretation of every other one, so the uncorrected column is retained rather than replaced: the models can perform these mappings when actually told what to produce, and the original ceiling was failing for the wrong reason.

Slot	System	cue in the sentence			no cue, no gold			
		<i>n</i>	commit	acc	<i>n</i>	commit	modal	share
rationality	Gemma 3 1B	228	75.9	88.4	144	89.6	UYAR	96.1
	Gemma 3 4B	228	89.5	100.0	144	97.9	UYAR	99.3
	Gemma 3 12B	228	86.4	100.0	144	99.3	UYAR	100.0
	Gemma 3 27B	228	86.0	100.0	144	100.0	UYAR	100.0
	Gemma 3n E2B	228	87.3	99.5	144	100.0	UYAR	100.0
	Gemma 3n E4B	228	90.3	99.5	144	100.0	UYAR	100.0
	Qwen3 8B	228	73.7	97.6	144	98.6	UYAR	97.9
	Qwen3 14B	228	79.4	98.9	144	99.3	UYAR	100.0
	Qwen3 32B	228	79.0	100.0	144	100.0	UYAR	100.0
	Llama 3.1 8B	228	74.1	97.6	144	93.1	UYAR	100.0
	Sarvam-Translate	228	76.8	99.4	144	98.6	UYAR	99.3
	NLLB-200 3.3B	228	80.3	92.3	144	100.0	UYAR	96.5
	IndicTrans2 1B	228	80.3	99.5	144	95.8	UYAR	98.6
gender	Gemma 3 1B	220	46.4	87.2	150	33.3	FEM	54.0
	Gemma 3 4B	220	64.1	79.4	150	60.0	MASC	72.2
	Gemma 3 12B	220	83.6	82.6	150	49.3	MASC	85.1
	Gemma 3 27B	220	82.7	84.6	150	48.7	MASC	91.8
	Gemma 3n E2B	220	69.5	85.6	150	32.7	MASC	67.3
	Gemma 3n E4B	220	69.5	89.5	150	32.7	MASC	83.7
	Qwen3 8B	220	77.7	81.9	150	43.3	FEM	52.3
	Qwen3 14B	220	70.5	89.7	150	28.0	<i>n/a</i>	<i>n/a</i>
	Qwen3 32B	220	43.2	87.4	150	14.0	<i>n/a</i>	<i>n/a</i>
	Llama 3.1 8B	220	40.9	94.4	150	3.3	<i>n/a</i>	<i>n/a</i>
	Sarvam-Translate	220	69.1	80.9	150	43.3	FEM	69.2
	NLLB-200 3.3B	220	25.0	<i>n/a</i>	150	16.7	<i>n/a</i>	<i>n/a</i>
	IndicTrans2 1B	220	27.7	<i>n/a</i>	150	17.3	<i>n/a</i>	<i>n/a</i>
number	Gemma 3 1B	112	58.0	69.2	68	36.8	<i>n/a</i>	<i>n/a</i>
	Gemma 3 4B	112	89.3	61.0	68	95.6	PL	92.3
	Gemma 3 12B	112	92.0	73.8	68	85.3	PL	75.9
	Gemma 3 27B	112	87.5	68.4	68	86.8	PL	74.6
	Gemma 3n E2B	112	90.2	62.4	68	73.5	PL	86.0
	Gemma 3n E4B	112	92.0	61.2	68	91.2	PL	96.8
	Qwen3 8B	112	79.5	59.6	68	97.1	PL	72.7
	Qwen3 14B	112	82.1	64.1	68	85.3	PL	93.1
	Qwen3 32B	112	83.9	51.1	68	97.1	PL	90.9
	Llama 3.1 8B	112	73.2	51.2	68	79.4	PL	96.3
	Sarvam-Translate	112	69.6	57.7	68	86.8	PL	83.0
	NLLB-200 3.3B	112	79.5	67.4	68	86.8	PL	100.0
	IndicTrans2 1B	112	79.5	56.2	68	80.9	PL	100.0

Table 9: **An in-sentence cue against no cue: the comparison the construction licenses.** The two conditions share a frame and a cue type and differ in whether the cue is present at all, so the pair isolates cue presence. **In gender the two cued conditions do not:** they cue the value through different channels, the prior-sentence condition scores higher on 11 of 13 systems, and no gender comparison between them is drawn (§8). The cue-free condition carries no gold by construction, so it has no accuracy; the emitted distribution is reported instead, as the modal value and the share of committed items taking it. A cell prints *n/a* below the reporting floor (< 30 committed items or $< 30\%$ commitment), because a ratio over a handful of items is not a rate. In gender, commitment falls when the cue is deleted on 13 of 13 systems, by 4.1–42.4 points. In rationality it falls on 0, and the systems answer a cue-free item as confidently as a cued one; in number it falls on 5, a slot whose results are not interpreted (§6).

Verb-form set	lexicon only (%)	+guesser (%)	+fallback (%)
UD_Tamil-TTB test (264 forms)	38.6	64	69.3
UD_Tamil-MWTT test (295 forms)	62	75.3	89.8
<i>Precision / recall, -ஆன் masculine slice on MWTT</i>			
precision (audited)			100
recall (audited)			98.7
recall, universal-after-context-filter			94.9
recall, existential (robustness)			94.9
<i>Slot-value extractor, benchmark gold</i>			
clusivity / gender / number / rationality recall	100.0 / 100.0 / 100.0 / 100.0		
honorificity recall			79.0
contrast false positives			0

Table 10: Analyser coverage as an ablation, and precision/recall on held-out UD_Tamil-MWTT. The fallback layer is reported because it was necessary, not because it flatters the number. The checker is *not* 100% accurate and we do not claim it is. Scoring follows the universal-after-context-filtering policy, with existential as a robustness column. Recall is reported for the withdrawn honorificity slot as well, because it is the measurement that withdrew it: 79.0% reflects the imperative sub-family, whose optative targets carry no exponent ThamizhiMorph will license (Appendix E).

Direct FEATS count	tokens	Consequence
Gender=Fem in UD_Tamil-TTB	0	no masc./fem. contrast is expressible
-ஆன்/-ஆள் finite verbs in TTB	0	the exponents themselves never occur
Clusivity across both treebanks	0	நாம்/நாங்கள் are not distinguished
-வில்லை/இல்லை tagged Polarity=Pos	41 / 41	the most frequent negation strategy is mislabelled
Aspect across both treebanks	0	unannotated
Deixis across both treebanks	0	documented as Tamil-specific, never used
Polite distinct values	1	one value only; the binary contrast is not annotated
Voice=Cau	0	causation unannotated
Mood ∈ {Pot,Nec,Des}	0	modality unannotated
Evident	0	evidentiality unannotated

Table 11: **Reproducible annotation facts**, counted directly from UD_Tamil-TTB (train/dev/test) and UD_Tamil-MWTT (test) by `scripts/ud_negative_facts.py`. Taken together, the standard Tamil annotation cannot express 7 categories a reference grammar treats as central: aspect, causation, clusivity, deixis, evidentiality, honorific contrast, modality. Tamil is also absent from UniMorph entirely. These are counts rather than arguments, which is why they open the paper. TTB is CC BY-NC-SA 3.0 and only counts are reproduced.