

Which Model Should Judge That Language?

A Controlled Multilingual Benchmark of Fluency Judges

François Remy^{1,2} Pierre-Carl Langlais^{1,3,4} Anastasia Stasenko^{1,5} Pieter Delobelle^{1,6}

¹Pleias

²Parallia ³Sorbonne Center for Artificial Intelligence ⁴Sciences Po Médialab

⁵Paris-Dauphine University ⁶KU Leuven

Abstract

Researchers increasingly use large language models as general-purpose evaluators, yet choosing a judge remains difficult when its competence, cost, and latency vary across tasks and languages. In this work, we study this problem for multilingual fluency judgment: assigning a score from 0 to 5 to indicate how natural a text sounds to a native speaker. In an initial pilot on grammatical error correction data, we found wide variation in how much teachers edit broken texts, making comparisons across languages difficult. We therefore introduce the Translationese Detection Benchmark (TDBENCH), built from original-language web texts and three synthetic levels of synthetic translationese: subtle, mild, and heavy. A judge succeeds when it scores the source above its intended degradation. Across 25 languages, detection depends strongly on the language and the strength of the degradation, suggesting that selecting the right judge is sensitive to the target language and the required sensitivity. Among the models we tested, Claude Fable 5.1 achieved the highest estimated macro detection rate (53.5%) albeit at a prohibitive cost. Gemini 3.7 Flash and GPT-6 Astra followed among general-purpose LLMs, scoring 48.7% and 45.8%. Remarkably, the specialized pretrained classifier Jev achieved 47.7% on the benchmark while costing significantly less. We detail these findings in this publication, and provide advice for judge evaluation.

1 Introduction

LLM-as-a-judge evaluation has become a practical default for tasks that exact-match metrics cannot capture (Liu et al., 2023; Zheng et al., 2023). Before trusting a model as a judge, we need to establish that it can recognize the differences in quality that matter for the task. Doing so is difficult because expert human judgments are expensive and noisy, while agreement with another automated evaluator can reproduce the same biases. Judge

selection is a model-selection problem in its own right, with quality, cost, latency, availability, and coverage as competing objectives.

We investigate this problem in the setting of multilingual fluency: given a text and its target language, a judge assigns a score between 0 and 5, where 0 denotes unintelligible text and 5 denotes fully natural text that could have been written by a native speaker. Measuring fluency matters when evaluating language models, translations, and synthetic data. We evaluate each model separately by language and corruption severity, thereby providing enough details for our readers to choose a judge appropriate for the language and fluency level they target.

We evaluate whether the scores put paired texts in the expected *order*. Starting from a source text, we construct increasingly translationese-like variants while preserving content. A competent judge should score the source above each degradation and should, ideally, order all four variants by severity. For example, both a drop from 5 to 4 and a drop from 3 to 2 count as successful detection. Figure 1 shows this procedure for one passage.

This article answers two questions: *How reliably can a judge detect controlled losses of naturalness across languages and disfluency levels?* and *Which judges remain practical choices when we account for monetary cost?*

We begin with a pilot on MultiGEC learner texts and teacher corrections, then introduce TDBENCH, which contains three levels of synthetic translationese for 1,250 passages written in 25 European languages. We compare several SOTA general-purpose LLMs with a specialized pretrained classifier as an alternative for routine fluency scoring.

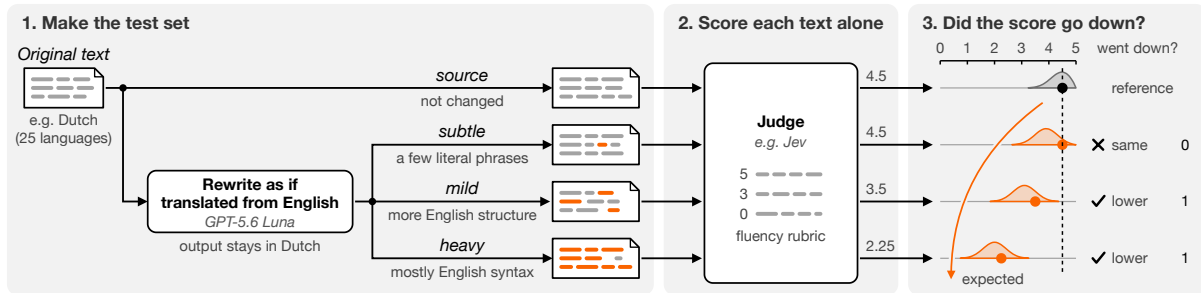


Figure 1: **How we test a fluency judge on one passage.** (1) We keep an original web passage as the source. GPT-5.6 Luna rewrites it in the same language as if it were translated from English, at three levels; orange marks English-like wording. (2) The judge scores each text alone and never sees two versions together. It uses a fluency rubric (0: not understandable, 3: clear but awkward, 5: native-like), adapted to its own scale. (3) Curves show where we expect the scores: lower with more English influence. Dots show the scores of the judge. A variant counts as detected only if its score is lower than the source; a tie or a higher score counts as a miss. The Dutch example and all scores are illustrative.

2 Background and Related Work

2.1 LLMs-as-a-judge and their known meta-evaluation pitfalls

LLM-as-a-judge evaluation became popular because strong models often agree reasonably well with human preferences on open-ended generation. G-Eval and MT-Bench helped establish this approach, while also exposing familiar weaknesses such as position bias, verbosity preference, and sensitivity to prompt wording (Liu et al., 2023; Zheng et al., 2023). Later work has shown that these effects can be large enough to change model rankings (Dubois et al., 2024; Shi et al., 2025).

Multilingual evaluation adds a failure mode: competence can vary by language. Hada et al. (2024b) find that LLM evaluators do not align equally well with human judgments across languages, especially outside high-resource settings. METAL reaches a similar conclusion using native-speaker judgments for multilingual summarization (Hada et al., 2024a). MM-Eval broadens the picture to 18 languages and shows that judge discrimination tends to weaken as language resources decrease (Son et al., 2024). Fu and Liu report low cross-language agreement across 25 languages and show that scaling the model or adding multilingual training does not automatically solve the problem (Fu and Liu, 2025).

These differences motivate validating evaluators separately in each language (Doğruöz et al., 2026). But building tests of comparable difficulty across languages remains a challenge. We try to resolve it using a proxy question: *can an LLM judge detect subtle losses of naturalness in otherwise under-*

standable text?

2.2 The challenge of measuring naturalness in practice

Most multilingual judge benchmarks mix several abilities: correctness, helpfulness, reasoning, safety, style, and language quality. MIST, for example, evaluates multilingual generation and includes human naturalness judgments, while MM-Eval combines linguistic and non-linguistic preference tasks (Kocmi et al., 2025; Son et al., 2024). Because they combine several criteria, their results leave sensitivity to small changes in naturalness difficult to assess separately.

A sentence can be grammatical yet still sound translated, overly literal, or lexically awkward. Conversely, a fluent sentence can be factually wrong or semantically inadequate. We therefore evaluate naturalness separately from factual accuracy and preservation of meaning.

Grammatical error correction corpora provided our first test cases. MultiGEC (Masciolini et al., 2025) contains learner texts paired with human corrections in multiple European languages. If a judge is sensitive to linguistic quality, it should generally prefer the corrected version. Controlled comparisons across MultiGEC languages are difficult because learner proficiency, error types, corpus sources, and the extent of teacher rewrites vary between datasets.

Translationese gives us a way to introduce subtle awkwardness into otherwise fluent text. Text translated from another language often carries traces of the source language in its syntax, lexical choices, and collocations (Ni et al., 2022). A recent English–

Swedish dataset asks models to distinguish idiomatic Swedish from translationese. They often prefer the more literal version even when they never see the English source (Kunz et al., 2026). Recognizing native-like phrasing thus poses a challenge even when the text is grammatical.

2.3 Controlled perturbations as a diagnostic tool for LLMs

Generating several versions of the same passage lets us change the wording while keeping the content approximately constant. We can then ask whether a judge’s score moves in the expected direction as the degradation becomes stronger.

Chang et al. (2025) find that multilingual evaluators are more aligned with human ratings in languages where they are also more sensitive to controlled perturbations. Perturbation sensitivity may thus serve as an inexpensive proxy for evaluator competence.

The proxy is imperfect: synthetic corruption can introduce artifacts, and a supposedly degraded sentence may occasionally match or improve on its source. MM-Eval explicitly warns that translated evaluation data can accidentally alter the intended preference label (Son et al., 2024). The Swedish translationese results make the same problem visible from another angle: even strong models can disagree with human notions of idiomaticity (Kunz et al., 2026).

Our subtle, mild, and heavy corruption levels describe the changes requested from the generator. The benchmark measures whether a judge responds to those changes as expected. We still need native speakers to assess how often the intended ordering matches their judgments.

2.4 Specialized approaches to multilingual evaluation

Specialized evaluators offer another route. M-Prometheus shows that smaller models trained specifically for multilingual evaluation can compete with much larger general-purpose judges (Pombal et al., 2025). CE-Judge similarly shows that a better evaluation procedure can improve multilingual judging without changing the underlying model (Ghiasvand Mohammadkhani and Beigy, 2025).

Choosing a judge for deployment requires evidence about its accuracy in the target language, its cost, and its response time. We compare hosted

LLMs with the specialized Jev classifier on that basis.

We extend this work by testing whether judge reliability changes with the *subtlety of the linguistic distinction*, using paired texts with approximately the same content.

3 Benchmark Development

3.1 Learner-written and teacher-corrected text

Our first iteration used learner texts and teacher corrections from MultiGEC (Masciolini et al., 2025). For each pair, we scored the learner sentence and its correction independently and counted the pair as correctly ordered when the correction received the higher score. The pair provides a human-authored comparison without requiring an absolute fluency label: the teacher judged the second sentence to improve on the first.

We sampled 600 pairs from ten languages, with 50 pairs for each of twelve combinations of language and edit style. We evaluated eight judges at this stage; their macro correction-reward accuracy ranged from 57.3% for Nemotron 3.5 Lightning to 84.3% for Gemini 3.7 Flash. MultiGEC also distinguishes minimal corrections from fluency-enhancing rewrites, a potentially useful test of sensitivity to small and large improvements. Averaging over the available language–style combinations, we found that Gemini ordered 86.0% of minimal pairs and 79.3% of fluency-enhancing pairs correctly. The two averages cover different sets of languages, so language differences may explain some of the gap.

Still, correction labels provide only limited insight into the difficulty of a comparison. While they describe how teachers edit, the size of the improvement varies within each category. Minimal corrections requires repairing errors that are immediately salient, but what that means precisely is naturally subjective. Besides, the broader rewrite of fluency edits may replace an acceptable sentence with a stylistic alternative whose superiority is not necessarily obvious. The learner populations, source domains, and annotation practices also differ across the component corpora. Even in Estonian and Ukrainian, the two languages for which our sample contains both edit styles, the size and direction of the style gap vary across judges. That being said, the pilot showed that judges usually reward human corrections. However, to study how accu-

racy changes with the size of the fluency difference, we had to build a second benchmark with three controlled levels of degradation for every source passage.

3.2 Controlled translationese construction

To achieve controllability, we built TD-BENCH from the original-language side of FineTranslations-Edu (Penedo et al., 2026). For each of 25 language-script groups, we sampled 50 documents with at least three aligned chunks and selected one interior chunk using a fixed random seed. The sample contains 1,250 source passages, equally divided among the languages. We use the original-language passage as the least translationese member of each set. This assumes that the sampled web passage was originally written in that language; individual passages may violate that assumption.

We then asked GPT-5.6 Luna, through OpenAI Codex in September 2026, to produce three increasingly translationese-like versions of each passage. We processed each passage independently, supplying the generator with the target-language identifier and source text. We considered also supplying the English translation distributed with FineTranslations, then judged it unnecessary given the model’s ability to translate the passage itself. The prompt asks the generator to preserve meaning, facts, names, numbers, paragraph boundaries, and approximate length while introducing progressively stronger English influence.

The *subtle* corruption level retains mostly natural target-language wording but permits a few literal calques and slightly stiff constructions. The *mild* corruption level follows English clause structure more closely and introduces several calques or awkward collocations. The *heavy* corruption level uses English syntax and discourse structure more consistently, together with frequent literal choices and occasional plausible transliteration errors for rare terms.

We paired each source passage with its three generated variants to test changes in fluency while holding content fixed. The complete benchmark contains 5,000 text positions, of which 4,985 are unique because a small number of passages recur in the source data. The severity labels reflect the requested changes; their agreement with native-speaker judgments remains to be tested.

4 Evaluation Protocol

4.1 Fluency scoring task

Each judge sees one passage at a time and rates its fluency and naturalness on a floating-point scale from 0 to 5. The lower endpoint denotes unintelligible text, 3 denotes understandable but noticeably awkward text, and 5 denotes text that reads like an original passage written by a fluent speaker. The prompt does not reveal whether the passage is the source or one of the three treatments, and no judge sees two versions of the same passage together. Pointwise presentation prevents a model from solving the task by directly comparing parallel strings.

The hosted generative models return a single score in increments of 0.25 and infer the language from the passage itself. Jev receives the language code and returns an expected score over a ten-level ordinal rubric, which we normalize to the same 0–5 range. M-Prometheus and the CE-Judge adaptation receive the language name and use integer ratings from 1 to 5, linearly mapped to 0–5. The judges may assign different meanings to the same number and allow different intervals between scores. We compare whether each judge lowers its score after a degradation. We include the full prompts and each system’s output format in the evaluation artifacts.

4.2 Paired metrics

For source row i , language λ , and degradation level $k \in \{\text{subtle}, \text{mild}, \text{heavy}\}$, let $s_{i,0}$ be the source score and $s_{i,k}$ the degraded score. We define detection as

$$d_{i,k} = \mathbb{K}[s_{i,0} > s_{i,k}], \quad (1)$$

and the paired score change as

$$\Delta_{i,k} = s_{i,k} - s_{i,0}. \quad (2)$$

Negative Δ indicates successful sensitivity to degradation. For each language and severity, the detection rate is the mean of $d_{i,k}$. Our primary score is the unweighted mean over the resulting language-severity cells, so each evaluated language and each treatment level contributes equally. We count a tie as zero detection and record cases where the judge scores the degraded passage higher as reversals. Perfect detection would yield 100%. A chance baseline would depend on how often a judge assigns equal scores, so 50% should not be interpreted as a common chance rate.

We report the mean score change Δ as a secondary measure because judges use the scale differently. For each set of four texts, we also check

whether scores decrease with severity, both allowing ties and requiring a strict decrease. Finally, we count the fraction of the six possible text pairs that the judge orders correctly. This gives credit for each correctly ordered pair even when the full sequence is wrong. For confidence intervals and model contrasts, we resample source passages within each language: all three comparisons from a passage remain together in the cluster bootstrap.

4.3 Comparison sets and missing scores

Because API reliability differs across models, we specify which rows each comparison includes. We retain a row only when all four scores for that model are available. If a model has an incomplete row in the comparison set, we treat all four scores as tied, giving zero detection credit for that row. We report the number of complete rows separately.

Our main tables use all 1,250 source rows for every model except Fable, retaining the additional observations available for those judges. For Fable, we estimate detection from the 1,015 rows it completed within the budget, about 81.2% of the dataset, and we exclude rows we didn’t send from this estimate. For paired model comparisons and language-level analysis, we also evaluate every judge on the same 1,015 rows, which we call the *Fable-supported set*.

5 Experimental Setup

We compare nine hosted general-purpose LLMs, the Jev pretrained classifier, and five locally run open-judge configurations. Figure 2 summarizes the macro detection rate of every judge; the released metadata provides exact model identifiers and complete-row coverage. We ran the hosted evaluations between September 4 and 18, 2026 through OpenAI-compatible Nebius and OpenRouter endpoints. We used OpenRouter for Fable 5.1, Gemini 3.7 Flash, and GPT Astra, and Nebius for Qwen3.5 and Nemotron Ultra. We set temperature to zero where the endpoint exposed it, and allowed up to three attempts after a failed request or unparsable answer. We sent the same prompt to all nine hosted models.

We evaluated Jev through TypeSafe’s System One score interface. We supplied a ten-level fluency rubric and normalized its probability-weighted ordinal output to 0–5. We selected this score formulation mostly because it is the interface’s native representation for ordinal judgments,

after a small pilot confirmed that the model performed similarly in several alternative designs.

We evaluate all three released M-Prometheus checkpoints, which were trained for multilingual direct assessment and pairwise comparison (Pommal et al., 2025). We retain the official integer 1–5 direct-assessment format, replace the general rubric with our fluency rubric, and map the result linearly to 0–5. We leave M-Prometheus’s reference-answer field empty, matching the information we give the other judges.

We further adapt the pointwise CE-Judge procedure (Ghiasvand Mohammadkhani and Beigy, 2025). Using its released Qwen2.5-7B-Instruct backbone, the pipeline first extracts relevant linguistic concepts, then turns them into passage-specific questions, and finally applies that checklist when scoring. The published prompts assume a source–translation pair; our *CE-Judge adapted* condition instead makes all three stages reference-free and specific to fluency. To separate the procedure from its backbone, we also evaluate Qwen2.5-7B-Instruct directly with the same 1–5 rubric.

We ran the five open-judge configurations in bfloat16 with greedy decoding on two RTX PRO 6000 Blackwell GPUs. We preserve failed requests and parsing failures in the run logs and charge all attempts, successful or not, when reporting API cost.

To contextualize language-level results, we construct a descriptive resource index from six public signals: estimated speakers, number of countries where the language is spoken, number of countries where it is official, Wikipedia article and active-editor counts, and the number of FineTranslations documents. We log-scale each signal to its 95th percentile across the metadata inventory, then average the six values. The index approximates how much digital data is available for each language. The amount of that data used to train each model is unknown. We include all 25 groups in aggregate results and discuss language-level differences among the 23 groups that the metadata represents as distinct languages.

6 Results

6.1 Which judge works best overall?

Table 1 compares all fifteen judges using full-benchmark results, except for Fable’s estimate which comes from its completed rows only. ▷

These results show that the benchmark is chal-

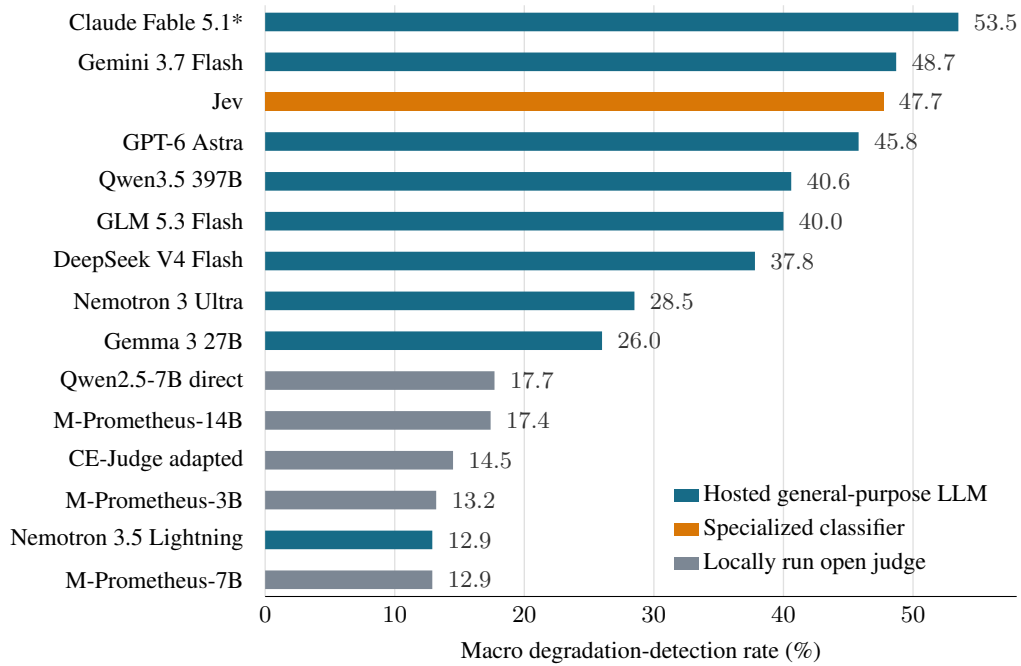


Figure 2: **Macro detection rates for all evaluated judges.** Bars are sorted by degradation-detection rate; colors distinguish hosted general-purpose LLMs, the specialized Jev classifier, and locally run open judges. Except for Fable, rates use all 1,250 source rows, with unavailable rows counted as ties. Fable’s starred estimate uses its 1,015 completed rows; counting its missing rows as ties gives 43.4%.

lenging but not impossible, and is not yet saturated. The fact the largest models perform better than small ones gives confidence the measured signal is real. Appendix A gives the complete results by language and severity, including coverage and score changes.

► **Fable leads the general-purpose LLMs, followed by Gemini and Astra.** Fable’s estimated macro detection rate is 53.5%, based on 81.2% of the dataset; Gemini scores 48.7% and Astra 45.8% on the full benchmark. On the same 1,015 passages completed by Fable, the three score 53.5%, 48.6%, and 45.5%, respectively. Paired bootstrap comparisons on this common set put Fable’s advantage over Gemini at 4.9 percentage points (95% CI: 3.5–6.4) and Gemini’s advantage over Astra at 3.1 points (1.5–4.6), using 10,000 source-row resamples within languages and seed 42. Gemini also leads Astra by 3.0 points on the full benchmark, with 1,249 complete rows compared with Astra’s 1,250. Jev sits between Gemini and Astra on the primary metric, scoring 47.7% on the full benchmark and 48.3% on the Fable-supported set.

► **Even the strongest judge detects only slightly more than half of the degradations.** On the full benchmark, the remaining hosted models score 40.0% for GLM 5.3 Flash, 37.8% for

DeepSeek V4 Flash, 28.5% for Nemotron 3 Ultra, 26.0% for Gemma 3 27B, and 12.9% for Nemotron 3.5 Lightning.

It is possible that Jev might score particularly high because it was not constrained by the 0.25 thresholds other LLMs targeted, resulting in fewer ties. Because the impact of thresholding depends on the absolute scale used by each model, we leave their analysis to future work.

► **The custom evaluators in Table 1 are considerably weaker on this task.** Direct Qwen2.5-7B-Instruct obtains 17.7% macro detection, and the strongest M-Prometheus checkpoint, 14B, obtains 17.4%. The adapted CE-Judge procedure reaches 14.5%, below its direct Qwen2.5-7B backbone. M-Prometheus-3B and 7B obtain 13.2% and 12.9%, respectively. Several of their bootstrap intervals overlap, leaving their relative ranking uncertain. All five configurations score well below the leading judges on this fluency task. Their integer scales also produce many ties: macro tie rates range from 62.4% for direct Qwen2.5-7B to 74.4% for M-Prometheus-7B, and ties count as non-detections under our metric.

6.2 Does corruption level matter?

► **Subtle translationese remains difficult even**

Judge	Detect $\uparrow$	Δ $\downarrow$	Heavy	Mild	Sub.	95% CI	N
Fable 5.1 [†]	53.5	-0.660	79.5	65.5	15.5	[51.6, 55.4]	1015
Gemini 3.7 Flash	48.7	-0.499	76.2	59.8	10.3	[47.2, 50.3]	1249
Jev	47.7	-0.050	74.1	50.0	18.9	[45.9, 49.4]	1250
GPT-6 Astra	45.8	-0.268	73.7	56.7	7.0	[44.2, 47.3]	1250
Qwen3.5 397B	40.6	-0.242	65.8	43.0	13.0	[38.8, 42.3]	1247
GLM 5.3 Flash	40.0	-0.143	67.0	45.0	8.2	[38.5, 41.6]	1250
DeepSeek V4 Flash	37.8	-0.231	64.0	38.1	11.3	[36.2, 39.5]	1250
Nemotron 3 Ultra	28.5	+0.038	33.8	28.1	23.7	[26.6, 30.4]	1146
Gemma 3 27B	26.0	-0.106	52.1	18.3	7.6	[24.5, 27.7]	1250
Qwen2.5-7B direct	17.7	+0.016	19.0	17.6	16.5	[16.1, 19.4]	1237
M-Prometheus-14B	17.4	-0.068	26.6	15.6	9.9	[15.8, 19.1]	1248
CE-Judge adapted	14.5	-0.000	17.9	13.8	11.8	[13.1, 15.9]	1249
M-Prometheus-3B	13.2	+0.225	21.4	11.5	6.6	[11.9, 14.5]	1245
Nemotron 3.5 Lightning	12.9	-0.041	15.2	12.4	11.2	[11.4, 14.5]	1249
M-Prometheus-7B	12.9	-0.006	17.8	11.4	9.5	[11.5, 14.4]	1249

Table 1: **Full-benchmark results with a partial-data estimate for Fable.** Macro and severity-specific degradation-detection rates (%). [†]Fable’s estimate uses only its 1,015 completed rows (81.2% of the dataset); all other scores use 1,250 rows, with unavailable rows counted as ties. Detection is in percent; Δ is the mean degraded-minus-source score (more negative means a larger drop), and N gives complete-row coverage. Bold and underlining mark the best and second-best distinct values in each score column; delta comparisons depend on the judge’s scoring scale. Confidence intervals use 2,000 source-row cluster-bootstrap samples within each model’s comparison set (seed 42). Colors follow the displayed detection rates: blue $\geq 45\%$, green $35\text{--}< 45\%$, amber $20\text{--}< 35\%$, and gray $< 20\%$. Rows use overall detection; severity cells use their own rates. Colors and highlights do not denote statistical significance.

for the leading judges. Fable detects 15.5% of subtle variants, compared with 65.5% of mild and 79.5% of heavy variants. The same gradient appears for the five leading judges in Table 1. Fable, Gemini, and Astra retain that order at every level on their common comparison set, so their overall ranking reflects differences across all three levels of corruption. Among those five systems, Jev is best on subtle changes at 18.9%, yet it trails Fable and Gemini on both mild and heavy changes.

Fable’s mean score change Δ is +0.226 for subtle translationese, -0.723 for mild, and -1.482 for heavy. Fable rates subtle variants above their sources on average, even though the requested intervention was a degradation. Some cases may be genuine misses; in others, the generator may have smoothed an imperfection already present in the web source. The average scores follow the requested severity levels. Native speakers would need to check whether that ordering holds for individual sets of texts.

On the common Fable-supported set, the ranking changes under the all-pairs metric. Fable leads on source-versus-degraded detection, whereas Jev slightly leads on accuracy over all six severity pairs, 67.9% to 67.4%. Jev’s nearly continuous scores produce far fewer ties. It can correctly order the degraded variants even when it fails to rank the original above them.

6.3 Does the best judge depend on the language?

For the language-level comparisons in this section, we use the common Fable-supported set so that each judge scores the same passages. Appendix A also reports every model’s full results, retaining Fable’s completed-row estimates. Language-level averages vary almost as much as severity-level averages. Across the 23 language groups considered in this analysis, Fable ranges from 63.6% for Spanish to 14.7% for Scottish Gaelic. French, Swedish, Italian, and Czech are also above 63%, while Galician, Irish, and Maltese are below 38%.

► **Languages with more resources generally receive higher detection rates.** Fable’s language ranking has a Spearman correlation of $\rho = 0.74$ with our descriptive resource index. For the other non-Nemotron general-purpose judges and Jev, the corresponding correlations range from 0.52 to 0.77. Departures from this trend may reflect the domains of the sampled passages, the corruption generator, differences in model training, or sampling variation.

► **The best judge depends on the language.** Jev, for example, reaches 65.9% on Lower Sorbian and 41.0% on Galician, exceeding Fable in both cases despite trailing it overall. Nemotron 3 Ultra has the highest observed rate on Maltese

at 40.8% and shares the highest rate on Scottish Gaelic at 23.3%, but reaches only 12.7% on Welsh. Nemotron 3.5 Lightning likewise has isolated relative successes on Lower Sorbian (30.1%) and Irish (26.7%), alongside misses on Icelandic (0.0%), Dutch (2.6%), and Czech (3.5%).

These Nemotron profiles are nearly uncorrelated with our resource index: $\rho = 0.01$ for Ultra and -0.19 for Lightning. The models’ documentation suggests one possible explanation. Although the base models see broader multilingual data, multilingual post-training is concentrated on a short list of languages, including French, German, Italian, Spanish, Chinese, and Japanese (NVIDIA, 2026a,b). Even that list does not predict the results cleanly: Lightning scores only 7.1% on German, 8.5% on Italian, and 10.1% on Spanish, while doing better on languages outside the stated post-training focus. These uneven results make direct testing in the target language necessary, including for languages listed as supported by the provider.

6.4 What does the best judge cost?

▷ **Fable’s lead came at a cost of roughly \$200 for 81.2% coverage.** That expenditure produced 1,015 complete rows. Including unsuccessful and retried requests, we spent roughly \$0.20 per completed source row, or \$49 per 1,000 completed text judgments.

Jev completed the full benchmark for 0.38\$ and scored 47.7%, compared with Fable’s 53.5% estimate. On their common comparison set, Jev scored slightly higher on all-pairs ordering. Its relative accuracy however varied by language, so replacing Fable with Jev requires checking performance in the intended setting. The result motivates using pretrained or custom-trained classifiers for routine scoring, potentially reserving expensive LLM calls for uncertain cases. We focus on LLM judges because they are widely used and because their ability to recognize natural phrasing may also tell us something about their own fluency when generating text (even though that possible link remains to be proved).

7 Discussion

We recommend validating a judge on the language and fluency differences it will encounter. Fable leads overall on the rows it completes, but Jev performs better on subtle changes and on some individual languages. Choosing by aggregate accuracy

alone could lead users to pay for a model whose advantage disappears on their task. A judge that detects obvious corruption may still struggle with the near-native cases that require careful human review.

Language resources help explain the broad pattern of results. However, the Nemotron models depart sharply from that pattern and produce language-level results too erratic to support claims based on isolated high ranks. Their selective multilingual post-training offers a possible explanation, although our experiment cannot identify which training choices caused these differences.

Paired comparisons make it possible to test judges even if they attach different meanings to a score of 4. Using those scores as absolute quality thresholds would require a separate calibration study with human ratings. A practical next step is to test a system that sends routine cases to a low-cost classifier and ambiguous ones to a more expensive general model. Because hosted endpoints and training mixtures change, each model version requires renewed validation.

Whether fluency judgment predicts generation ability remains an open question. A direct test would pair the present judge scores with native-speaker ratings of text generated by the same models in each language. Such an experiment could establish whether the ability to recognize translationese predicts the ability to avoid producing it.

8 Limitations

The benchmark’s central limitation is that its ordering is generated rather than human-validated. A shared source passage keeps content approximately constant, but the three severity levels come from instructions to a single generator. A variant may alter the intended meaning, introduce a conspicuous artifact unrelated to translationese, or accidentally improve an awkward source. Judges may also respond to generator-specific wording, limiting generalization to other forms of translationese. Future versions should therefore use multiple corruption models and ask native speakers to verify meaning preservation and to rerank the texts.

Generalization is further constrained by the source data and language coverage. FineTranslations passages may themselves be translated, non-native, code-switched, formulaic, or poorly written, while our English-centered corruption prompt does not model interference from other source lan-

guages. Moreover, the 25 language–script groups are predominantly European and contain only 50 source passages each. That being said, limiting the evaluation to the 42 Fable-supported rows returned very similar percentages, showing the risk is limited. Still, broader claims about multilingual fluency judging would require more languages, domains, and sources of linguistic interference.

Some measured differences may also reflect scoring interfaces rather than fluency competence alone. Hosted LLMs use half-point ratings, Jev returns a near-continuous expected ordinal score, and M-Prometheus and CE-Judge use coarser integer ratings. Coarser scales produce more ties and thus lower measured detection, even though the metric considers only the direction of score changes. Our CE-Judge adaptation also departs from the published task by removing the source text used in translation assessment. Alternative prompts and direct pairwise comparisons would help determine how strongly these design choices affect the results.

Missing scores create an additional source of uncertainty. Fable and Nemotron Ultra have many incomplete rows, and their API failures may depend on language or passage length. We report both accuracy on the Fable-supported set and results that count missing scores as ties, but estimating complete-run performance still requires assumptions about why requests failed. Hosted model identifiers may also drift behind an endpoint, limiting exact reproducibility even when prompts and caches are released.

Finally, we have not yet measured native-speaker performance on the benchmark. The current scores therefore measure agreement with the ordering requested from the generator rather than agreement with human judgments. We deduplicate the fifteen repeated positions among 5,000 at request time and keep each source row’s three variants together when estimating uncertainty, but the language-level comparisons remain exploratory. Larger samples, native-speaker validation, and correction for multiple comparisons would allow more rigorous tests of the observed language differences.

9 Conclusion

We introduced TDBENCH, a controlled multilingual benchmark that tests whether a candidate judge scores an original-language passage above increasingly translationese-like variants. The paired design lets us compare judges by whether their scores decrease after a requested loss of fluency.

Among the general-purpose LLMs we tested, Fable 5.1 has the highest estimated detection rate, followed by Gemini 3.7 Flash and GPT-6 Astra. Comparisons on their common set of passages support both gaps. However, the leading model changes with language and corruption severity, and subtle translationese remains difficult across the comparison. At approximately \$200 for 81.2% coverage, Fable is expensive for routine use; Jev achieves competitive accuracy and completes the full benchmark. The open multilingual judge methods tested here perform poorly on this task of scoring passages in isolation.

Future work should validate the generated ordering with native speakers, broaden the range of languages, varieties, domains, and source-language interference, and test whether judging competence predicts generation fluency. We will release the benchmark data, prompts, score caches, and evaluation scripts to support language-specific, cost-aware choices of fluency judge.

Ethical Considerations

FineTranslations derives from public web data under the ODC-By 1.0 license and Common Crawl terms. Released artifacts should retain provenance, respect upstream removals, and avoid presenting passages as newly authored; synthetic variants may preserve identifying details from their sources.

Naturalness judgments can privilege standardized varieties and penalize legitimate dialectal, learner, or code-switched usage. Native-speaker validation should therefore represent relevant varieties, document annotator backgrounds, and compensate linguistic expertise. We also report coverage, retries, hardware, and cost to make the financial and computational trade-offs of large judges visible.

References

- Jiayi Chang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025. [Exploring the multilingual NLG evaluation abilities of LLM-based evaluators](#). *Preprint*, arXiv:2503.04360.
- A. Seza Doğruöz, Xixian Liao, Verena Blaschke, Jakob Prange, Senyu Li, and David Ifeoluwa Adelani. 2026. [Challenges and recommendations for LLMs-as-a-judge in multilingual settings and low-resource languages](#). *Preprint*, arXiv:2607.02235.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. [Length-controlled alpaca-eval: A simple way to debias automatic evaluators](#). *Preprint*, arXiv:2404.04475. Published at COLM 2024.
- Xiyan Fu and Wei Liu. 2025. [How reliable is multilingual LLM-as-a-judge?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.
- Mohammad Ghiasvand Mohammadkhani and Hamid Beigy. 2025. [Checklist engineering empowers multilingual LLM judges](#). In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 190–196, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.
- Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. [METAL: Towards multilingual meta-evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.
- Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. [Are large language model-based evaluators the solution to scaling up multilingual evaluation?](#) In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian’s, Malta. Association for Computational Linguistics.
- Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova, and Vilém Zouhar. 2025. [Findings of the WMT25 multilingual instruction shared task: Persistent hurdles in reasoning, generation, and evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 414–435, Suzhou, China. Association for Computational Linguistics.
- Jenny Kunz, Anja Jarochochenko, and Marcel Bollmann. 2026. [A dataset for probing translationese preferences in English-to-Swedish translation](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 8767–8779, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using GPT-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfah, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. [The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL](#). In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–33, Tallinn, Estonia. University of Tartu Library.
- Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. [Original or translated? a causal analysis of the impact of translationese on machine translation performance](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.
- NVIDIA. 2026a. NVIDIA Nemotron 3 Ultra 550B A55B model card. <https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16>.
- NVIDIA. 2026b. NVIDIA Nemotron 3.5 Lightning 30B A3B model card. <https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-BF16>.
- Guilherme Penedo, Hynek Kydlíček, Amir Hossein Karagan, and Leandro von Werra. 2026. FineTranslations. <https://huggingface.co/datasets/HuggingFaceFW/finetranslations>.
- José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. [M-prometheus: A suite of open multilingual LLM judges](#). In *Conference on Language Modeling (COLM)*. Also available as arXiv:2504.04953.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. [Judging the judges: A systematic study of position bias in LLM-as-a-judge](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aulabasco, Mano Aslan, Vu Trong Kim, Shayekh Bin

Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. [MM-eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models](#). *Preprint*, arXiv:2410.17578.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Complete Model Comparison

We include all numeric breakdowns from `results/translationese/model_comparison.html`, replacing Fable's entries with its completed-row estimates from the companion Fable-supported report. As in Table 1, Fable scores 53.5% based on 1,015 completed source rows (81.2% coverage). All other models use the full set of 1,250 rows, with unavailable rows counted as ties. For each language, Fable's rates use only the passages it completed; the detailed tables give that denominator explicitly. Comparisons with Fable therefore involve different passage sets, and its estimates may be affected by which requests failed. Colors use the same detection thresholds throughout: **blue** $\geq 45\%$, **green** $35\% < 45\%$, **amber** $20\% < 35\%$, and **gray** $< 20\%$. Each cell or row takes its color from the detection rate it reports, so a model can change color across languages and severity levels. Bold marks the best value and underlining the second-best distinct value in each comparison; tied values share the highlight. Higher detection is better; a more negative score change Δ means a larger drop after corruption. Delta highlights describe score sensitivity and depend on the judge's scale; they do not establish a ranking of judge quality. The colors and highlights are descriptive and do not denote statistical significance. Every table uses the same model order: descending average detection across all languages and severity levels, including Fable's completed-row estimate. The appendix includes all 25 language-script groups in the released benchmark. Language-specific results describe these samples of 50 passages and should guide further validation in the intended use case.

A.1 Model identifiers and aggregate results

Tables use model-name abbreviations: FAB (Fable), GEM (Gemini), JEV (Jev), GPT (GPT-6 Astra), Q35 (Qwen3.5), GLM, DSK (DeepSeek), NMU (Nemotron Ultra), GMA (Gemma), Q25 (Qwen2.5), MP14/MP3/MP7 (M-Prometheus by size), CEJ (CE-Judge), and NML (Nemotron Lightning).

Table 2: Model key and complete-row coverage out of 1,250.

Code	Model identifier	Scored rows	Missing
FAB	anthropic/claude-fable-5.1	1015	235
GEM	gemini-3.7-flash	1249	1
JEV	jev/jev-latest:score	1250	0
GPT	gpt-6-astra	1250	0
Q35	Qwen/Qwen3.5-397B-A17B	1247	3
GLM	zai-org/GLM-5.3-Flash	1250	0
DSK	deepseek-ai/DeepSeek-V4-Flash-0731	1250	0
NMU	nvidia/Nemotron-3-Ultra-550b-a55b	1146	104
GMA	google/gemma-3-27b-it	1250	0
Q25	Qwen/Qwen2.5-7B-Instruct	1237	13
MP14	Unbabel/M-Prometheus-14B	1248	2
CEJ	CE-Judgeadapted/Qwen/Qwen2.5-7B-Instruct	1249	1
MP3	Unbabel/M-Prometheus-3B	1245	5
NML	nvidia/Nemotron-3.5-Lightning	1249	1
MP7	Unbabel/M-Prometheus-7B	1249	1

Detection averages the rate of scoring the source above its degradation across language-severity cells. Mean Δ is the degraded score minus the source score, with equal weight for each language-severity cell. It is the negative of the report's discrimination metric. Monotonic measures the share of source-subtle-mild-heavy sequences whose scores never increase; strict monotonic requires a decrease at every step. Ordered pairs measures accuracy over all six pairs in each sequence. For models other than Fable, missing rows contribute zero discrimination and zero strict-order accuracy, and satisfy the non-strict monotonic condition because all four scores are tied. We exclude Fable's missing rows from all its metrics.

Table 3: Model ranking and ordering metrics, with completed-row estimates for Fable. Rates are percentages; Δ is in score units.

Model	Detect $\uparrow$	Δ $\downarrow$	Ordered pairs	Monotonic	Strict	Scored rows
FAB	53.5	-0.660	67.4	44.3	4.7	1015
GEM	<u>48.7</u>	-0.499	65.2	35.0	2.4	1249
JEV	47.7	-0.050	67.7	10.5	9.4	1250

Table 3 (continued)

Model	Detect $\uparrow$	Δ $\downarrow$	Ordered pairs	Monotonic	Strict	Scored rows
GPT	45.8	-0.268	64.0	40.2	2.6	1250
Q35	40.6	-0.242	56.7	32.2	2.7	1247
GLM	40.0	-0.143	58.4	31.6	2.3	1250
DSK	37.8	-0.231	51.9	38.9	0.8	1250
NMU	28.5	+0.038	33.0	32.9	0.6	1146
GMA	26.0	-0.106	38.9	57.7	0.1	1250
Q25	17.7	+0.016	18.0	53.6	0.1	1237
MP14	17.4	-0.068	22.2	65.1	0.0	1248
CEJ	14.5	-0.000	15.3	64.1	0.0	1249
MP3	13.2	+0.225	17.8	66.4	0.0	1245
NML	12.9	-0.041	13.7	73.0	0.0	1249
MP7	12.9	-0.006	14.0	71.9	0.1	1249

A.2 Which models perform best in each language?

Table 4 shows detection by language, grouped into heavy, mild, and subtle translationese blocks separated by horizontal rules. All fifteen models appear in the same order throughout, sorted by their overall average detection score. Bold marks every model tied for the highest unrounded rate within a language and severity; underlining marks the next distinct rate. The highlights describe observed rankings and do not establish statistically significant differences.

Table 4: Detection by language (%), grouped by translationese level: heavy, mild, then subtle. Models follow overall average detection order; the mean row in each block averages equally over languages. Fable uses completed passages only. Codes refer to Table 2.

Language	FAB	GEM	JEV	GPT	Q35	GLM	DSK	NMU	GMA	Q25	MP14	CEJ	MP3	NML	MP7
Heavy translationese															
Basque	63.8	66.0	52.0	60.0	42.0	36.0	34.0	20.0	26.0	36.0	8.0	14.0	6.0	14.0	6.0
Czech	97.4	<u>98.0</u>	100.0	<u>98.0</u>	88.0	92.0	94.0	28.0	76.0	10.0	24.0	6.0	28.0	10.0	34.0
Dutch	<u>97.4</u>	96.0	98.0	94.0	86.0	92.0	86.0	28.0	74.0	14.0	32.0	16.0	22.0	14.0	28.0
Estonian	81.4	76.0	<u>80.0</u>	74.0	76.0	70.0	72.0	30.0	52.0	24.0	22.0	8.0	28.0	6.0	8.0
French	<u>95.1</u>	96.0	96.0	96.0	92.0	88.0	94.0	50.0	86.0	20.0	42.0	24.0	52.0	34.0	36.0
Galician	<u>51.3</u>	44.0	54.0	40.0	38.0	44.0	32.0	20.0	22.0	10.0	12.0	18.0	8.0	10.0	10.0
German	97.6	92.0	90.0	<u>94.0</u>	86.0	88.0	84.0	46.0	84.0	10.0	20.0	20.0	22.0	12.0	20.0
Greek	100.0	<u>98.0</u>	92.0	96.0	90.0	94.0	86.0	64.0	52.0	24.0	24.0	10.0	12.0	4.0	12.0
Hungarian	<u>87.5</u>	92.0	92.0	92.0	80.0	86.0	76.0	26.0	64.0	16.0	32.0	6.0	28.0	10.0	0.0
Icelandic	88.1	86.0	78.0	<u>88.0</u>	78.0	78.0	64.0	20.0	38.0	16.0	18.0	12.0	18.0	0.0	10.0
Irish	42.5	<u>34.0</u>	<u>34.0</u>	24.0	24.0	22.0	24.0	16.0	16.0	30.0	28.0	32.0	10.0	28.0	18.0
Italian	<u>97.4</u>	98.0	92.0	98.0	88.0	92.0	88.0	48.0	82.0	18.0	50.0	22.0	46.0	12.0	26.0
Latvian	87.2	<u>88.0</u>	<u>88.0</u>	92.0	64.0	80.0	66.0	18.0	74.0	20.0	20.0	18.0	16.0	6.0	20.0
Limburgish	75.0	<u>68.0</u>	52.0	60.0	60.0	42.0	50.0	32.0	20.0	32.0	34.0	24.0	12.0	30.0	18.0
Lower Sorbian	<u>75.6</u>	48.0	76.0	58.0	48.0	56.0	62.0	44.0	34.0	14.0	24.0	18.0	12.0	34.0	20.0
Luxembourgish	73.8	<u>64.0</u>	58.0	50.0	42.0	<u>34.0</u>	42.0	50.0	34.0	30.0	36.0	28.0	20.0	28.0	18.0
Maltese	52.5	<u>52.0</u>	44.0	<u>52.0</u>	<u>34.0</u>	36.0	34.0	34.0	26.0	26.0	28.0	16.0	18.0	4.0	14.0
Portuguese	89.3	<u>86.0</u>	80.0	<u>86.0</u>	72.0	78.0	70.0	50.0	62.0	8.0	20.0	36.0	32.0	22.0	18.0
Scottish Gaelic	14.0	<u>28.0</u>	30.0	16.0	26.0	22.0	20.0	20.0	10.0	<u>28.0</u>	14.0	24.0	12.0	14.0	8.0
Slovenian	92.9	<u>92.0</u>	<u>92.0</u>	90.0	80.0	86.0	74.0	22.0	66.0	12.0	22.0	12.0	18.0	6.0	6.0
Spanish	<u>97.7</u>	98.0	98.0	98.0	78.0	88.0	92.0	46.0	84.0	6.0	40.0	20.0	44.0	18.0	34.0
Swedish	95.0	<u>90.0</u>	88.0	<u>90.0</u>	76.0	80.0	78.0	28.0	76.0	2.0	34.0	12.0	12.0	18.0	34.0
Ukrainian	97.8	94.0	<u>96.0</u>	<u>96.0</u>	92.0	<u>96.0</u>	88.0	46.0	84.0	16.0	20.0	14.0	36.0	2.0	16.0
Welsh	71.4	<u>56.0</u>	50.0	<u>50.0</u>	54.0	50.0	54.0	12.0	36.0	28.0	30.0	28.0	12.0	20.0	18.0
Western Frisian	65.9	<u>64.0</u>	42.0	50.0	50.0	44.0	36.0	48.0	24.0	26.0	30.0	10.0	12.0	24.0	12.0
Mean	79.5	<u>76.2</u>	74.1	73.7	65.8	67.0	64.0	33.8	52.1	19.0	26.6	17.9	21.4	15.2	17.8
Mild translationese															
Basque	57.4	<u>48.0</u>	44.0	<u>48.0</u>	24.0	36.0	32.0	20.0	18.0	32.0	12.0	8.0	4.0	8.0	8.0
Czech	84.2	78.0	72.0	<u>84.0</u>	52.0	74.0	50.0	20.0	26.0	8.0	6.0	8.0	14.0	8.0	14.0
Dutch	78.9	70.0	66.0	64.0	48.0	54.0	48.0	24.0	14.0	4.0	22.0	14.0	6.0	8.0	14.0
Estonian	<u>62.8</u>	66.0	46.0	66.0	48.0	44.0	38.0	24.0	12.0	22.0	12.0	14.0	18.0	4.0	10.0
French	85.4	<u>78.0</u>	72.0	<u>78.0</u>	66.0	70.0	56.0	24.0	24.0	14.0	12.0	10.0	18.0	30.0	10.0
Galician	<u>33.3</u>	32.0	48.0	18.0	30.0	22.0	20.0	20.0	16.0	16.0	16.0	10.0	6.0	10.0	12.0
German	76.2	68.0	66.0	<u>74.0</u>	62.0	66.0	62.0	20.0	26.0	10.0	6.0	16.0	4.0	8.0	8.0
Greek	<u>79.5</u>	76.0	54.0	82.0	52.0	60.0	32.0	36.0	12.0	16.0	4.0	12.0	4.0	2.0	4.0
Hungarian	<u>77.5</u>	68.0	54.0	80.0	38.0	38.0	22.0	24.0	10.0	12.0	14.0	4.0	16.0	4.0	4.0
Icelandic	66.7	62.0	26.0	<u>66.0</u>	34.0	36.0	24.0	28.0	4.0	14.0	8.0	8.0	6.0	0.0	10.0
Irish	45.0	30.0	24.0	22.0	18.0	14.0	12.0	16.0	12.0	<u>40.0</u>	26.0	36.0	8.0	26.0	22.0
Italian	87.2	<u>86.0</u>	70.0	84.0	58.0	72.0	50.0	30.0	18.0	12.0	14.0	8.0	18.0	8.0	10.0
Latvian	<u>53.8</u>	54.0	44.0	50.0	36.0	50.0	38.0	42.0	10.0	8.0	8.0	18.0	8.0	8.0	14.0
Limburgish	<u>63.9</u>	74.0	36.0	56.0	38.0	40.0	52.0	32.0	14.0	26.0	26.0	26.0	16.0	24.0	14.0
Lower Sorbian	<u>63.4</u>	50.0	64.0	52.0	48.0	44.0	60.0	54.0	30.0	12.0	30.0	20.0	6.0	28.0	14.0
Luxembourgish	69.0	<u>64.0</u>	44.0	42.0	48.0	12.0	36.0	52.0	34.0	34.0	26.0	20.0	20.0	14.0	16.0
Maltese	37.5	36.0	38.0	34.0	32.0	<u>40.0</u>	18.0	42.0	22.0	22.0	28.0	16.0	18.0	14.0	8.0
Portuguese	75.0	<u>54.0</u>	44.0	52.0	38.0	50.0	30.0	24.0	18.0	6.0	12.0	12.0	12.0	18.0	6.0
Scottish Gaelic	18.6	26.0	<u>30.0</u>	14.0	20.0	16.0	16.0	22.0	8.0	36.0	14.0	14.0	12.0	4.0	6.0
Slovenian	64.3	62.0	54.0	<u>64.0</u>	42.0	42.0	34.0	28.0	10.0	8.0	0.0	10.0	0.0	4.0	12.0
Spanish	88.4	78.0	68.0	<u>86.0</u>	54.0	66.0	54.0	22.0	34.0	0.0	18.0	10.0	18.0	14.0	10.0
Swedish	80.0	<u>70.0</u>	56.0	<u>70.0</u>	50.0	60.0	42.0	16.0	22.0	4.0	12.0	16.0	14.0	12.0	20.0
Ukrainian	73.3	68.0	58.0	<u>70.0</u>	60.0	<u>70.0</u>	54.0	28.0	14.0	20.0	12.0	4.0	10.0	4.0	6.0
Welsh	59.5	<u>52.0</u>	44.0	34.0	44.0	28.0	40.0	12.0	28.0	24.0	32.0	16.0	20.0	18.0	24.0
Western Frisian	56.1	<u>44.0</u>	28.0	28.0	34.0	20.0	32.0	42.0	22.0	40.0	20.0	16.0	12.0	24.0	10.0
Mean	65.5	<u>59.8</u>	50.0	56.7	43.0	45.0	38.1	28.1	18.3	17.6	15.6	13.8	11.5	12.4	11.4
Subtle translationese															
Basque	25.5	16.0	30.0	22.0	18.0	14.0	12.0	18.0	14.0	<u>26.0</u>	6.0	4.0	2.0	10.0	4.0
Czech	7.9	4.0	<u>16.0</u>	2.0	12.0	6.0	6.0	34.0	2.0	8.0	0.0	8.0	2.0	6.0	12.0
Dutch	7.9	4.0	<u>18.0</u>	4.0	4.0	4.0	4.0	24.0	2.0	10.0	2.0	10.0	0.0	6.0	8.0
Estonian	4.7	6.0	18.0	6.0	2.0	6.0	2.0	<u>22.0</u>	0.0	24.0	10.0	20.0	4.0	6.0	10.0
French	9.8	4.0	26.0	8.0	10.0	<u>12.0</u>	8.0	10.0	4.0	<u>12.0</u>	4.0	4.0	4.0	8.0	0.0
Galician	5.1	6.0	14.0	0.0	<u>12.0</u>	2.0	10.0	14.0	6.0	14.0	10.0	6.0	6.0	6.0	10.0
German	7.1	4.0	16.0	2.0	2.0	2.0	2.0	4.0	2.0	<u>8.0</u>	2.0	6.0	0.0	2.0	<u>8.0</u>
Greek	4.5	2.0	8.0	2.0	10.0	6.0	2.0	24.0	2.0	<u>12.0</u>	2.0	2.0	0.0	2.0	0.0
Hungarian	<u>12.5</u>	4.0	10.0	8.0	4.0	4.0	0.0	24.0	4.0	8.0	4.0	2.0	6.0	12.0	6.0
Icelandic	9.5	6.0	<u>14.0</u>	10.0	6.0	10.0	2.0	24.0	0.0	<u>14.0</u>	8.0	<u>14.0</u>	8.0	0.0	10.0
Irish	<u>22.5</u>	20.0	24.0	4.0	12.0	2.0	8.0	14.0	12.0	30.0	<u>28.0</u>	30.0	12.0	24.0	16.0
Italian	5.1	4.0	10.0	4.0	<u>12.0</u>	4.0	2.0	16.0	2.0	10.0	2.0	<u>12.0</u>	2.0	6.0	10.0
Latvian	5.1	4.0	<u>14.0</u>	2.0	8.0	6.0	10.0	30.0	2.0	<u>14.0</u>	0.0	10.0	2.0	4.0	<u>14.0</u>

Table 4 (continued)

Language	FAB	GEM	JEV	GPT	Q35	GLM	DSK	NMU	GMA	Q25	MP14	CEJ	MP3	NML	MP7
Limburgish	52.8	36.0	16.0	20.0	34.0	22.0	30.0	<u>38.0</u>	10.0	18.0	24.0	18.0	12.0	26.0	10.0
Lower Sorbian	<u>34.1</u>	26.0	<u>48.0</u>	28.0	32.0	28.0	56.0	<u>46.0</u>	30.0	12.0	20.0	8.0	8.0	30.0	12.0
Luxembourgish	50.0	30.0	32.0	12.0	30.0	4.0	30.0	50.0	30.0	<u>36.0</u>	34.0	28.0	20.0	22.0	16.0
Maltese	22.5	24.0	<u>26.0</u>	18.0	12.0	18.0	16.0	40.0	14.0	24.0	16.0	8.0	16.0	14.0	10.0
Portuguese	0.0	0.0	4.0	0.0	4.0	4.0	6.0	<u>12.0</u>	0.0	10.0	2.0	20.0	6.0	6.0	2.0
Scottish Gaelic	11.6	2.0	12.0	2.0	8.0	4.0	10.0	<u>22.0</u>	4.0	38.0	16.0	18.0	8.0	16.0	10.0
Slovenian	4.8	0.0	<u>18.0</u>	2.0	12.0	6.0	2.0	38.0	2.0	4.0	2.0	8.0	4.0	8.0	8.0
Spanish	4.7	<u>8.0</u>	10.0	2.0	6.0	<u>8.0</u>	2.0	10.0	0.0	4.0	4.0	4.0	2.0	<u>8.0</u>	4.0
Swedish	15.0	4.0	20.0	2.0	10.0	4.0	10.0	<u>16.0</u>	4.0	8.0	0.0	10.0	4.0	6.0	10.0
Ukrainian	4.4	4.0	<u>12.0</u>	0.0	6.0	4.0	6.0	18.0	0.0	<u>12.0</u>	4.0	4.0	6.0	8.0	8.0
Welsh	35.7	22.0	44.0	12.0	<u>38.0</u>	16.0	24.0	10.0	24.0	<u>24.0</u>	30.0	24.0	14.0	20.0	24.0
Western Frisian	<u>24.4</u>	18.0	12.0	2.0	<u>22.0</u>	8.0	22.0	34.0	20.0	<u>32.0</u>	18.0	16.0	16.0	24.0	16.0
Mean	15.5	10.3	<u>18.9</u>	7.0	13.0	8.2	11.3	23.7	7.6	16.5	9.9	11.8	6.6	11.2	9.5

A.3 Complete results by language and severity

The following tables include all 1,125 model–language–severity combinations, with completed-row estimates for Fable. Each table starts with detection ($\uparrow$) and mean score change Δ ($\downarrow$), followed by uncertainty, coverage, and secondary diagnostics. We highlight detection and Δ independently within each language–severity block. Scored/ N gives the number of complete rows and the denominator used for detection. Mean score uses only scored rows, while mean Δ (degraded minus source), detection, ties, and reversals include missing rows as ties with zero change for models other than Fable. For Fable, both counts equal the number of completed passages and vary by language; every other model uses $N = 50$. Detection, ties, and reversals sum to 100%, apart from rounding. The 95% intervals here describe mean Δ and use the report's normal approximation (mean ± 1.96 standard errors). They differ from the bootstrap intervals for detection in the main table. Each language has its own table, with heavy, mild, and subtle blocks separated by horizontal rules. Within every block, models retain the overall average detection ranking.

A.3.1 Basque

Table 5: Basque: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	63.8	-0.457	[-0.796, -0.119]	47/47	10.6	25.5	3.213
GEM	66.0	-0.570	[-0.890, -0.250]	50/50	12.0	22.0	3.460
JEV	52.0	+0.030	[-0.194, +0.254]	50/50	0.0	48.0	3.751
GPT	60.0	-0.380	[-0.623, -0.137]	50/50	24.0	16.0	3.530
Q35	42.0	-0.100	[-0.384, +0.184]	49/50	22.0	36.0	3.541
GLM	36.0	-0.210	[-0.474, +0.054]	50/50	36.0	28.0	3.580
DSK	34.0	-0.010	[-0.347, +0.327]	50/50	36.0	30.0	4.070
NMU	20.0	+0.000	[-0.194, +0.194]	24/50	58.0	22.0	1.438
GMA	26.0	-0.010	[-0.174, +0.154]	50/50	50.0	24.0	4.050
Q25	36.0	-0.225	[-0.809, +0.359]	50/50	36.0	28.0	2.550
MP14	8.0	+0.700	[+0.180, +1.220]	50/50	64.0	28.0	3.925
CEJ	14.0	+0.325	[-0.070, +0.720]	50/50	68.0	18.0	4.750
MP3	6.0	+0.850	[+0.301, +1.399]	50/50	64.0	30.0	3.300
NML	14.0	+0.020	[-0.259, +0.299]	50/50	66.0	20.0	2.560
MP7	6.0	+0.275	[-0.097, +0.647]	50/50	80.0	14.0	4.650
Mild translationese							
FAB	57.4	-0.298	[-0.637, +0.042]	47/47	14.9	27.7	3.372
GEM	48.0	-0.190	[-0.494, +0.114]	50/50	20.0	32.0	3.840
JEV	44.0	+0.149	[-0.074, +0.372]	50/50	0.0	56.0	3.870
GPT	48.0	-0.100	[-0.347, +0.147]	50/50	30.0	22.0	3.810
Q35	24.0	+0.150	[-0.122, +0.422]	49/50	30.0	46.0	3.796
GLM	36.0	+0.120	[-0.148, +0.388]	50/50	18.0	46.0	3.910
DSK	32.0	+0.100	[-0.289, +0.489]	50/50	32.0	36.0	4.180
NMU	20.0	-0.060	[-0.249, +0.129]	24/50	62.0	18.0	1.312
GMA	18.0	+0.130	[-0.020, +0.280]	50/50	46.0	36.0	4.190
Q25	32.0	+0.150	[-0.447, +0.747]	50/50	36.0	32.0	2.925
MP14	12.0	+0.800	[+0.237, +1.363]	50/50	54.0	34.0	4.025
CEJ	8.0	+0.325	[+0.038, +0.612]	50/50	72.0	20.0	4.750
MP3	4.0	+1.000	[+0.481, +1.519]	50/50	62.0	34.0	3.450
NML	8.0	+0.180	[-0.102, +0.462]	50/50	70.0	22.0	2.720
MP7	8.0	+0.175	[-0.215, +0.565]	50/50	78.0	14.0	4.550
Subtle translationese							
FAB	25.5	+0.330	[-0.024, +0.684]	47/47	25.5	48.9	4.000
GEM	16.0	+0.460	[+0.190, +0.730]	50/50	34.0	50.0	4.490
JEV	30.0	+0.282	[+0.055, +0.509]	50/50	0.0	70.0	4.003
GPT	22.0	+0.320	[+0.088, +0.552]	50/50	28.0	50.0	4.230
Q35	18.0	+0.460	[+0.168, +0.752]	49/50	22.0	60.0	4.112
GLM	14.0	+0.530	[+0.287, +0.773]	50/50	24.0	62.0	4.320
DSK	12.0	+0.420	[+0.021, +0.819]	50/50	42.0	46.0	4.500
NMU	18.0	-0.060	[-0.304, +0.184]	24/50	68.0	14.0	1.312
GMA	14.0	+0.190	[+0.027, +0.353]	50/50	54.0	32.0	4.250
Q25	26.0	+0.075	[-0.517, +0.667]	50/50	38.0	36.0	2.850
MP14	6.0	+0.900	[+0.415, +1.385]	50/50	62.0	32.0	4.125
CEJ	4.0	+0.350	[+0.037, +0.663]	50/50	78.0	18.0	4.775
MP3	2.0	+0.875	[+0.355, +1.395]	50/50	68.0	30.0	3.325
NML	10.0	+0.330	[+0.003, +0.657]	50/50	66.0	24.0	2.870
MP7	4.0	+0.300	[-0.025, +0.625]	50/50	80.0	16.0	4.675

A.3.2 Czech

Table 6: Czech: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	97.4	-2.303	$[-2.532, -2.073]$	38/38	2.6	0.0	1.776
GEM	98.0	-2.150	$[-2.371, -1.929]$	50/50	2.0	0.0	2.030
JEV	100.0	-1.061	$[-1.183, -0.939]$	50/50	0.0	0.0	2.966
GPT	98.0	-1.600	$[-1.758, -1.442]$	50/50	2.0	0.0	2.530
Q35	88.0	-1.350	$[-1.573, -1.127]$	50/50	10.0	2.0	2.760
GLM	92.0	-1.440	$[-1.645, -1.235]$	50/50	6.0	2.0	2.510
DSK	94.0	-1.700	$[-1.941, -1.459]$	50/50	4.0	2.0	2.530
NMU	28.0	+0.060	$[-0.175, +0.295]$	45/50	38.0	34.0	1.100
GMA	76.0	-0.610	$[-0.736, -0.484]$	50/50	24.0	0.0	3.640
Q25	10.0	+0.025	$[-0.139, +0.189]$	50/50	78.0	12.0	4.225
MP14	24.0	-0.675	$[-1.066, -0.284]$	50/50	74.0	2.0	1.250
CEJ	6.0	+0.100	$[-0.054, +0.254]$	50/50	80.0	14.0	4.925
MP3	28.0	-0.575	$[-0.919, -0.231]$	50/50	68.0	4.0	1.475
NML	10.0	-0.110	$[-0.216, -0.004]$	50/50	90.0	0.0	2.980
MP7	34.0	-0.700	$[-1.170, -0.230]$	50/50	56.0	10.0	3.700
Mild translationese							
FAB	84.2	-0.987	$[-1.236, -0.738]$	38/38	5.3	10.5	3.092
GEM	78.0	-0.880	$[-1.092, -0.668]$	50/50	10.0	12.0	3.300
JEV	72.0	-0.083	$[-0.180, +0.015]$	50/50	0.0	28.0	3.945
GPT	84.0	-0.780	$[-0.944, -0.616]$	50/50	10.0	6.0	3.350
Q35	52.0	-0.430	$[-0.636, -0.224]$	50/50	30.0	18.0	3.680
GLM	74.0	-0.590	$[-0.791, -0.389]$	50/50	10.0	16.0	3.360
DSK	50.0	-0.370	$[-0.665, -0.075]$	50/50	26.0	24.0	3.860
NMU	20.0	+0.340	$[-0.057, +0.623]$	45/50	24.0	56.0	1.411
GMA	26.0	-0.050	$[-0.144, +0.044]$	50/50	60.0	14.0	4.200
Q25	8.0	+0.000	$[-0.140, +0.140]$	50/50	84.0	8.0	4.200
MP14	6.0	-0.175	$[-0.437, +0.087]$	50/50	90.0	4.0	1.750
CEJ	8.0	+0.050	$[-0.106, +0.206]$	50/50	80.0	12.0	4.875
MP3	14.0	-0.275	$[-0.566, +0.016]$	50/50	80.0	6.0	1.775
NML	8.0	-0.090	$[-0.190, +0.010]$	50/50	92.0	0.0	3.000
MP7	14.0	+0.050	$[-0.238, +0.338]$	50/50	72.0	14.0	4.450
Subtle translationese							
FAB	7.9	+0.382	$[-0.207, +0.557]$	38/38	36.8	55.3	4.461
GEM	4.0	+0.420	$[-0.274, +0.566]$	50/50	38.0	58.0	4.600
JEV	16.0	+0.243	$[-0.156, +0.329]$	50/50	0.0	84.0	4.270
GPT	2.0	+0.290	$[-0.185, +0.395]$	50/50	50.0	48.0	4.420
Q35	12.0	+0.360	$[-0.207, +0.513]$	50/50	26.0	62.0	4.470
GLM	6.0	+0.400	$[-0.247, +0.553]$	50/50	36.0	58.0	4.350
DSK	6.0	+0.510	$[-0.280, +0.740]$	50/50	52.0	42.0	4.740
NMU	34.0	+0.050	$[-0.191, +0.291]$	45/50	32.0	34.0	1.089
GMA	2.0	+0.170	$[-0.088, +0.252]$	50/50	66.0	32.0	4.420
Q25	8.0	+0.000	$[-0.140, +0.140]$	50/50	84.0	8.0	4.200
MP14	0.0	+0.400	$[-0.125, +0.675]$	50/50	82.0	18.0	2.325
CEJ	8.0	+0.050	$[-0.106, +0.206]$	50/50	80.0	12.0	4.875
MP3	2.0	+0.175	$[-0.057, +0.407]$	50/50	82.0	16.0	2.225
NML	6.0	+0.160	$[-0.018, +0.338]$	50/50	82.0	12.0	3.250
MP7	12.0	+0.150	$[-0.136, +0.436]$	50/50	72.0	16.0	4.550

A.3.3 Dutch

Table 7: Dutch: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	<u>97.4</u>	-1.987	[-2.239, -1.735]	38/38	2.6	0.0	2.000
GEM	96.0	-1.830	[-2.046, -1.614]	50/50	4.0	0.0	2.240
JEV	98.0	-0.684	[-0.820, -0.547]	50/50	0.0	2.0	3.267
GPT	94.0	-1.220	[-1.388, -1.052]	50/50	6.0	0.0	2.790
Q35	86.0	-1.190	[-1.412, -0.968]	50/50	12.0	2.0	2.780
GLM	92.0	-1.360	[-1.570, -1.150]	50/50	6.0	2.0	2.610
DSK	86.0	-1.560	[-1.821, -1.299]	50/50	14.0	0.0	2.600
NMU	28.0	-0.130	[-0.330, +0.070]	36/50	52.0	20.0	1.972
GMA	74.0	-0.510	[-0.617, -0.403]	50/50	26.0	0.0	3.660
Q25	14.0	+0.075	[-0.264, +0.414]	50/50	66.0	20.0	3.475
MP14	32.0	-0.600	[-0.923, -0.277]	50/50	64.0	4.0	1.525
CEJ	16.0	-0.125	[-0.313, +0.063]	50/50	76.0	8.0	4.600
MP3	22.0	-0.375	[-0.698, -0.052]	50/50	76.0	2.0	1.200
NML	14.0	-0.090	[-0.162, -0.018]	50/50	86.0	0.0	2.990
MP7	28.0	-0.250	[-0.563, +0.063]	50/50	56.0	16.0	3.675
Mild translationese							
FAB	78.9	-0.776	[-1.009, -0.543]	38/38	10.5	10.5	3.211
GEM	<u>70.0</u>	-0.700	[-0.934, -0.466]	50/50	18.0	12.0	3.370
JEV	66.0	+0.023	[-0.068, +0.114]	50/50	2.0	32.0	3.974
GPT	64.0	-0.530	[-0.692, -0.368]	50/50	30.0	6.0	3.480
Q35	48.0	-0.370	[-0.551, -0.189]	50/50	40.0	12.0	3.600
GLM	54.0	-0.500	[-0.700, -0.300]	50/50	34.0	12.0	3.470
DSK	48.0	-0.460	[-0.709, -0.211]	50/50	36.0	16.0	3.700
NMU	24.0	+0.080	[-0.188, +0.348]	36/50	50.0	26.0	2.264
GMA	14.0	-0.010	[-0.081, +0.061]	50/50	74.0	12.0	4.160
Q25	4.0	+0.275	[+0.086, +0.464]	50/50	72.0	24.0	3.675
MP14	22.0	+0.050	[-0.270, +0.370]	50/50	56.0	22.0	2.175
CEJ	14.0	+0.000	[-0.185, +0.185]	50/50	72.0	14.0	4.725
MP3	6.0	+0.050	[-0.263, +0.363]	50/50	84.0	10.0	1.625
NML	8.0	+0.060	[-0.076, +0.196]	50/50	82.0	10.0	3.140
MP7	14.0	+0.000	[-0.242, +0.242]	50/50	74.0	12.0	3.925
Subtle translationese							
FAB	7.9	+0.316	[+0.133, +0.498]	38/38	47.4	44.7	4.303
GEM	4.0	+0.580	[+0.418, +0.742]	50/50	26.0	70.0	4.650
JEV	<u>18.0</u>	+0.266	[+0.177, +0.356]	50/50	0.0	82.0	4.217
GPT	4.0	+0.400	[+0.288, +0.512]	50/50	32.0	64.0	4.410
Q35	4.0	+0.490	[+0.341, +0.639]	50/50	34.0	62.0	4.460
GLM	4.0	+0.370	[+0.210, +0.530]	50/50	54.0	42.0	4.340
DSK	4.0	+0.530	[+0.337, +0.723]	50/50	46.0	50.0	4.690
NMU	24.0	+0.040	[-0.221, +0.301]	36/50	46.0	30.0	2.208
GMA	2.0	+0.210	[+0.121, +0.299]	50/50	60.0	38.0	4.380
Q25	10.0	+0.125	[-0.088, +0.338]	50/50	74.0	16.0	3.525
MP14	2.0	+0.675	[+0.353, +0.997]	50/50	62.0	36.0	2.800
CEJ	10.0	+0.050	[-0.121, +0.221]	50/50	76.0	14.0	4.775
MP3	0.0	+0.550	[+0.199, +0.901]	50/50	82.0	18.0	2.125
NML	6.0	+0.190	[+0.045, +0.335]	50/50	74.0	20.0	3.270
MP7	8.0	+0.075	[-0.151, +0.301]	50/50	80.0	12.0	4.000

A.3.4 Estonian

Table 8: Estonian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	81.4	-1.663	$[-1.990, -1.335]$	43/43	11.6	7.0	1.977
GEM	76.0	-1.550	$[-1.882, -1.218]$	50/50	18.0	6.0	2.130
JEV	<u>80.0</u>	-0.402	$[-0.579, -0.225]$	50/50	0.0	20.0	3.213
GPT	74.0	-1.150	$[-1.393, -0.907]$	50/50	20.0	6.0	2.640
Q35	76.0	-1.130	$[-1.434, -0.826]$	50/50	12.0	12.0	2.730
GLM	70.0	-0.930	$[-1.194, -0.666]$	50/50	20.0	10.0	2.600
DSK	72.0	-1.260	$[-1.610, -0.910]$	50/50	10.0	18.0	2.580
NMU	30.0	+0.030	$[-0.272, +0.332]$	46/50	36.0	34.0	1.902
GMA	52.0	-0.260	$[-0.395, -0.125]$	50/50	32.0	16.0	3.760
Q25	24.0	+0.125	$[-0.266, +0.516]$	50/50	42.0	34.0	3.925
MP14	22.0	-0.375	$[-0.773, +0.023]$	50/50	66.0	12.0	1.700
CEJ	8.0	+0.025	$[-0.123, +0.173]$	50/50	82.0	10.0	4.800
MP3	28.0	-0.500	$[-0.975, -0.025]$	50/50	64.0	8.0	2.000
NML	6.0	+0.070	$[-0.088, +0.228]$	50/50	86.0	8.0	2.970
MP7	8.0	+0.250	$[-0.107, +0.607]$	50/50	74.0	18.0	4.550
Mild translationese							
FAB	62.8	-0.442	$[-0.732, -0.152]$	43/43	11.6	25.6	3.198
GEM	66.0	-0.400	$[-0.674, -0.126]$	50/50	2.0	32.0	3.280
JEV	46.0	+0.314	$[+0.126, +0.502]$	50/50	0.0	54.0	3.930
GPT	66.0	-0.290	$[-0.516, -0.064]$	50/50	8.0	26.0	3.500
Q35	48.0	-0.130	$[-0.359, +0.099]$	50/50	22.0	30.0	3.730
GLM	44.0	-0.070	$[-0.303, +0.163]$	50/50	24.0	32.0	3.460
DSK	38.0	-0.090	$[-0.338, +0.158]$	50/50	28.0	34.0	3.750
NMU	24.0	+0.210	$[-0.072, +0.492]$	46/50	34.0	42.0	2.098
GMA	12.0	+0.180	$[+0.052, +0.308]$	50/50	56.0	32.0	4.200
Q25	22.0	-0.125	$[-0.448, +0.198]$	50/50	54.0	24.0	3.675
MP14	12.0	-0.075	$[-0.399, +0.249]$	50/50	74.0	14.0	2.000
CEJ	14.0	-0.025	$[-0.203, +0.153]$	50/50	74.0	12.0	4.750
MP3	18.0	+0.175	$[-0.358, +0.708]$	50/50	62.0	20.0	2.675
NML	4.0	+0.130	$[-0.025, +0.285]$	50/50	86.0	10.0	3.030
MP7	10.0	+0.250	$[-0.107, +0.607]$	50/50	72.0	18.0	4.550
Subtle translationese							
FAB	4.7	+0.744	$[+0.419, +1.069]$	43/43	51.2	44.2	4.384
GEM	6.0	+0.830	$[+0.548, +1.112]$	50/50	36.0	58.0	4.510
JEV	18.0	+0.548	$[+0.334, +0.761]$	50/50	0.0	82.0	4.163
GPT	6.0	+0.580	$[+0.355, +0.805]$	50/50	42.0	52.0	4.370
Q35	2.0	+0.720	$[+0.451, +0.989]$	50/50	42.0	56.0	4.580
GLM	6.0	+0.840	$[+0.574, +1.106]$	50/50	32.0	62.0	4.370
DSK	2.0	+0.790	$[+0.496, +1.084]$	50/50	48.0	50.0	4.630
NMU	<u>22.0</u>	+0.280	$[+0.022, +0.538]$	46/50	32.0	46.0	2.174
GMA	0.0	+0.350	$[+0.209, +0.491]$	50/50	60.0	40.0	4.370
Q25	24.0	<u>-0.050</u>	$[-0.385, +0.285]$	50/50	52.0	24.0	3.750
MP14	10.0	+0.375	$[-0.052, +0.802]$	50/50	62.0	28.0	2.450
CEJ	20.0	-0.150	$[-0.330, +0.030]$	50/50	72.0	8.0	4.625
MP3	4.0	+0.575	$[+0.131, +1.019]$	50/50	76.0	20.0	3.075
NML	6.0	+0.040	$[-0.150, +0.230]$	50/50	86.0	8.0	2.940
MP7	10.0	+0.300	$[-0.032, +0.632]$	50/50	74.0	16.0	4.600

A.3.5 French

Table 9: French: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	<u>95.1</u>	-1.976	$[-2.236, -1.715]$	41/41	4.9	0.0	1.866
GEM	96.0	-1.850	$[-2.094, -1.606]$	50/50	2.0	2.0	2.190
JEV	96.0	-1.018	$[-1.196, -0.840]$	50/50	0.0	4.0	3.032
GPT	96.0	-1.280	$[-1.462, -1.098]$	50/50	2.0	2.0	2.670
Q35	92.0	-1.210	$[-1.408, -1.012]$	50/50	8.0	0.0	2.710
GLM	88.0	-1.280	$[-1.526, -1.034]$	50/50	6.0	6.0	2.510
DSK	94.0	-1.610	$[-1.861, -1.359]$	50/50	4.0	2.0	2.450
NMU	50.0	-0.450	$[-0.683, -0.217]$	48/50	42.0	8.0	2.635
GMA	86.0	-0.610	$[-0.723, -0.497]$	50/50	12.0	2.0	3.410
Q25	20.0	-0.250	$[-0.512, +0.012]$	50/50	74.0	6.0	3.175
MP14	42.0	-1.125	$[-1.569, -0.681]$	50/50	56.0	2.0	1.175
CEJ	24.0	-0.300	$[-0.593, -0.007]$	50/50	66.0	10.0	4.350
MP3	52.0	-1.100	$[-1.535, -0.665]$	50/50	44.0	4.0	0.825
NML	34.0	-0.460	$[-0.667, -0.253]$	50/50	66.0	0.0	2.940
MP7	36.0	-0.875	$[-1.273, -0.477]$	50/50	60.0	4.0	2.950
Mild translationese							
FAB	85.4	-1.110	$[-1.353, -0.867]$	41/41	9.8	4.9	2.732
GEM	<u>78.0</u>	-0.890	$[-1.133, -0.647]$	50/50	14.0	8.0	3.150
JEV	72.0	-0.112	$[-0.219, -0.005]$	50/50	0.0	28.0	3.938
GPT	<u>78.0</u>	-0.740	$[-0.913, -0.567]$	50/50	18.0	4.0	3.210
Q35	66.0	-0.560	$[-0.738, -0.382]$	50/50	26.0	8.0	3.360
GLM	70.0	-0.710	$[-0.948, -0.472]$	50/50	22.0	8.0	3.080
DSK	56.0	-0.610	$[-0.846, -0.374]$	50/50	36.0	8.0	3.450
NMU	24.0	-0.070	$[-0.313, +0.173]$	48/50	52.0	24.0	3.031
GMA	24.0	-0.050	$[-0.144, +0.044]$	50/50	60.0	16.0	3.970
Q25	14.0	-0.100	$[-0.330, +0.130]$	50/50	80.0	6.0	3.325
MP14	12.0	-0.075	$[-0.375, +0.225]$	50/50	80.0	8.0	2.225
CEJ	10.0	-0.100	$[-0.283, +0.083]$	50/50	84.0	6.0	4.550
MP3	18.0	-0.275	$[-0.540, -0.010]$	50/50	76.0	6.0	1.650
NML	30.0	-0.330	$[-0.541, -0.119]$	50/50	62.0	8.0	3.070
MP7	10.0	-0.100	$[-0.360, +0.160]$	50/50	84.0	6.0	3.725
Subtle translationese							
FAB	9.8	+0.488	$[+0.262, +0.713]$	41/41	34.1	56.1	4.329
GEM	4.0	+0.520	$[+0.322, +0.718]$	50/50	42.0	54.0	4.560
JEV	26.0	+0.261	$[+0.163, +0.359]$	50/50	0.0	74.0	4.311
GPT	8.0	+0.340	$[+0.196, +0.484]$	50/50	36.0	56.0	4.290
Q35	10.0	+0.460	$[+0.288, +0.632]$	50/50	32.0	58.0	4.380
GLM	<u>12.0</u>	+0.550	$[+0.340, +0.760]$	50/50	24.0	64.0	4.340
DSK	8.0	+0.520	$[+0.300, +0.740]$	50/50	50.0	42.0	4.580
NMU	10.0	+0.470	$[+0.220, +0.720]$	48/50	44.0	46.0	3.594
GMA	4.0	+0.270	$[+0.172, +0.368]$	50/50	46.0	50.0	4.290
Q25	<u>12.0</u>	+0.000	$[-0.210, +0.210]$	50/50	76.0	12.0	3.425
MP14	4.0	+0.275	$[+0.020, +0.530]$	50/50	80.0	16.0	2.575
CEJ	4.0	+0.050	$[-0.070, +0.170]$	50/50	88.0	8.0	4.700
MP3	4.0	+0.300	$[+0.015, +0.585]$	50/50	82.0	14.0	2.225
NML	8.0	+0.150	$[-0.034, +0.334]$	50/50	66.0	26.0	3.550
MP7	0.0	+0.450	$[+0.190, +0.710]$	50/50	78.0	22.0	4.275

A.3.6 Galician

Table 10: Galician: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	<u>51.3</u>	-0.423	[-0.723, -0.123]	39/39	35.9	12.8	2.897
GEM	44.0	-0.370	[-0.727, -0.013]	50/50	24.0	32.0	3.290
JEV	54.0	-0.003	[-0.115, +0.110]	50/50	0.0	46.0	3.638
GPT	40.0	-0.110	[-0.267, +0.047]	50/50	34.0	26.0	3.610
Q35	38.0	-0.320	[-0.679, +0.039]	50/50	32.0	30.0	3.420
GLM	44.0	-0.340	[-0.646, -0.034]	50/50	30.0	26.0	3.310
DSK	32.0	-0.300	[-0.669, +0.069]	50/50	30.0	38.0	3.420
NMU	20.0	+0.030	[-0.180, +0.240]	47/50	48.0	32.0	2.734
GMA	22.0	-0.130	[-0.304, +0.044]	50/50	62.0	16.0	4.110
Q25	10.0	-0.050	[-0.292, +0.192]	50/50	80.0	10.0	3.175
MP14	12.0	-0.050	[-0.321, +0.221]	50/50	80.0	8.0	1.950
CEJ	18.0	-0.050	[-0.330, +0.230]	50/50	64.0	18.0	4.550
MP3	8.0	+0.400	[+0.033, +0.767]	49/50	74.0	18.0	1.990
NML	10.0	-0.060	[-0.215, +0.095]	50/50	84.0	6.0	2.920
MP7	10.0	+0.100	[-0.179, +0.379]	50/50	74.0	16.0	3.950
Mild translationese							
FAB	33.3	-0.064	[-0.390, +0.262]	39/39	41.0	25.6	3.256
GEM	32.0	-0.030	[-0.392, +0.332]	50/50	24.0	44.0	3.630
JEV	48.0	+0.086	[-0.017, +0.189]	50/50	0.0	52.0	3.727
GPT	18.0	+0.140	[+0.009, +0.271]	50/50	44.0	38.0	3.860
Q35	30.0	-0.160	[-0.542, +0.222]	50/50	26.0	44.0	3.580
GLM	22.0	+0.090	[-0.160, +0.340]	50/50	46.0	32.0	3.740
DSK	20.0	+0.190	[-0.156, +0.536]	50/50	30.0	50.0	3.910
NMU	20.0	+0.180	[-0.072, +0.432]	47/50	40.0	40.0	2.894
GMA	16.0	-0.080	[-0.256, +0.096]	50/50	70.0	14.0	4.160
Q25	16.0	-0.025	[-0.318, +0.268]	50/50	68.0	16.0	3.200
MP14	16.0	+0.025	[-0.284, +0.334]	50/50	68.0	16.0	2.025
CEJ	10.0	+0.050	[-0.192, +0.292]	50/50	76.0	14.0	4.650
MP3	6.0	+0.400	[+0.040, +0.760]	49/50	78.0	16.0	1.990
NML	10.0	-0.050	[-0.207, +0.107]	50/50	86.0	4.0	2.930
MP7	12.0	-0.150	[-0.409, +0.109]	50/50	82.0	6.0	3.700
Subtle translationese							
FAB	5.1	+0.564	[+0.206, +0.923]	39/39	10.3	84.6	3.885
GEM	6.0	+0.630	[+0.244, +1.016]	50/50	18.0	76.0	4.290
JEV	14.0	+0.298	[+0.174, +0.422]	50/50	0.0	86.0	3.939
GPT	0.0	+0.630	[+0.499, +0.761]	50/50	20.0	80.0	4.350
Q35	<u>12.0</u>	+0.400	[+0.000, +0.800]	50/50	16.0	72.0	4.140
GLM	2.0	+0.590	[+0.401, +0.779]	50/50	44.0	54.0	4.240
DSK	10.0	+0.650	[+0.257, +1.043]	50/50	30.0	60.0	4.370
NMU	14.0	+0.300	[+0.048, +0.552]	47/50	40.0	46.0	3.021
GMA	6.0	+0.060	[-0.141, +0.261]	50/50	60.0	34.0	4.300
Q25	14.0	-0.100	[-0.370, +0.170]	50/50	74.0	12.0	3.125
MP14	10.0	+0.225	[-0.093, +0.543]	50/50	70.0	20.0	2.225
CEJ	6.0	+0.225	[-0.034, +0.484]	50/50	76.0	18.0	4.825
MP3	6.0	+0.825	[+0.368, +1.282]	49/50	66.0	28.0	2.423
NML	6.0	-0.070	[-0.226, +0.086]	50/50	90.0	4.0	2.910
MP7	10.0	+0.100	[-0.227, +0.427]	50/50	72.0	18.0	3.950

A.3.7 German

Table 11: German: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	97.6	-2.024	$[-2.250, -1.797]$	42/42	0.0	2.4	1.857
GEM	92.0	-1.650	$[-1.896, -1.404]$	50/50	6.0	2.0	2.190
JEV	90.0	-1.013	$[-1.203, -0.824]$	50/50	2.0	8.0	2.794
GPT	<u>94.0</u>	-1.300	$[-1.490, -1.110]$	50/50	4.0	2.0	2.630
Q35	86.0	-1.150	$[-1.376, -0.924]$	50/50	6.0	8.0	2.690
GLM	88.0	-1.340	$[-1.567, -1.113]$	50/50	8.0	4.0	2.420
DSK	84.0	-1.270	$[-1.530, -1.010]$	50/50	12.0	4.0	2.480
NMU	46.0	-0.330	$[-0.523, -0.137]$	47/50	40.0	14.0	2.511
GMA	84.0	-0.660	$[-0.793, -0.527]$	50/50	14.0	2.0	3.380
Q25	10.0	+0.125	$[-0.099, +0.349]$	50/50	74.0	16.0	3.875
MP14	20.0	-0.550	$[-0.908, -0.192]$	50/50	78.0	2.0	1.225
CEJ	20.0	-0.150	$[-0.399, +0.099]$	50/50	66.0	14.0	4.500
MP3	22.0	-0.350	$[-0.593, -0.107]$	50/50	76.0	2.0	0.850
NML	12.0	-0.060	$[-0.187, +0.067]$	50/50	84.0	4.0	2.920
MP7	20.0	-0.225	$[-0.558, +0.108]$	50/50	68.0	12.0	3.900
Mild translationese							
FAB	76.2	-0.964	$[-1.208, -0.720]$	42/42	16.7	7.1	2.917
GEM	68.0	-0.590	$[-0.848, -0.332]$	50/50	14.0	18.0	3.250
JEV	66.0	-0.029	$[-0.194, +0.137]$	50/50	0.0	34.0	3.779
GPT	<u>74.0</u>	-0.670	$[-0.857, -0.483]$	50/50	16.0	10.0	3.260
Q35	62.0	-0.400	$[-0.644, -0.156]$	50/50	18.0	20.0	3.440
GLM	66.0	-0.590	$[-0.799, -0.381]$	50/50	22.0	12.0	3.170
DSK	62.0	-0.430	$[-0.710, -0.150]$	50/50	22.0	16.0	3.320
NMU	20.0	+0.050	$[-0.154, +0.254]$	47/50	50.0	30.0	2.915
GMA	26.0	+0.000	$[-0.137, +0.137]$	50/50	52.0	22.0	4.040
Q25	10.0	+0.125	$[-0.063, +0.313]$	50/50	70.0	20.0	3.875
MP14	6.0	+0.125	$[-0.099, +0.349]$	50/50	74.0	20.0	1.900
CEJ	16.0	-0.025	$[-0.241, +0.191]$	50/50	68.0	16.0	4.625
MP3	4.0	+0.125	$[-0.120, +0.370]$	50/50	82.0	14.0	1.325
NML	8.0	+0.030	$[-0.129, +0.189]$	50/50	84.0	8.0	3.010
MP7	8.0	+0.125	$[-0.120, +0.370]$	50/50	82.0	10.0	4.250
Subtle translationese							
FAB	7.1	+0.476	$[+0.284, +0.668]$	42/42	35.7	57.1	4.357
GEM	4.0	+0.710	$[+0.489, +0.931]$	50/50	30.0	66.0	4.550
JEV	16.0	+0.401	$[+0.223, +0.578]$	50/50	0.0	84.0	4.208
GPT	2.0	+0.460	$[+0.297, +0.623]$	50/50	44.0	54.0	4.390
Q35	2.0	+0.760	$[+0.560, +0.960]$	50/50	26.0	72.0	4.600
GLM	2.0	+0.640	$[+0.472, +0.808]$	50/50	26.0	72.0	4.400
DSK	2.0	+0.820	$[+0.578, +1.062]$	50/50	32.0	66.0	4.570
NMU	4.0	+0.740	$[+0.468, +1.012]$	47/50	38.0	58.0	3.649
GMA	2.0	+0.270	$[+0.144, +0.396]$	50/50	60.0	38.0	4.310
Q25	<u>8.0</u>	<u>+0.225</u>	$[-0.004, +0.454]$	50/50	70.0	22.0	3.975
MP14	2.0	+0.950	$[+0.538, +1.362]$	50/50	60.0	38.0	2.725
CEJ	6.0	+0.175	$[+0.003, +0.347]$	50/50	74.0	20.0	4.825
MP3	0.0	+0.850	$[+0.451, +1.249]$	50/50	66.0	34.0	2.050
NML	2.0	+0.250	$[+0.062, +0.438]$	50/50	76.0	22.0	3.230
MP7	<u>8.0</u>	+0.275	$[-0.024, +0.574]$	50/50	76.0	16.0	4.400

A.3.8 Greek

Table 12: Greek: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	100.0	<u>-1.966</u>	[-2.185, -1.747]	44/44	0.0	0.0	2.000
GEM	98.0	-2.040	[-2.242, -1.838]	50/50	2.0	0.0	2.170
JEV	92.0	-0.608	[-0.730, -0.485]	50/50	2.0	6.0	3.384
GPT	96.0	-1.450	[-1.599, -1.301]	50/50	4.0	0.0	2.710
Q35	90.0	-1.260	[-1.450, -1.070]	50/50	10.0	0.0	2.810
GLM	94.0	-1.300	[-1.463, -1.137]	50/50	6.0	0.0	2.600
DSK	86.0	-1.420	[-1.692, -1.148]	50/50	12.0	2.0	2.640
NMU	64.0	-0.490	[-0.723, -0.257]	49/50	22.0	14.0	2.041
GMA	52.0	-0.320	[-0.428, -0.212]	50/50	46.0	2.0	3.720
Q25	24.0	-0.125	[-0.407, +0.157]	50/50	62.0	14.0	3.725
MP14	24.0	-0.425	[-0.702, -0.148]	49/50	74.0	2.0	1.352
CEJ	10.0	-0.100	[-0.254, +0.054]	50/50	86.0	4.0	4.675
MP3	12.0	-0.175	[-0.583, +0.233]	50/50	78.0	10.0	1.925
NML	4.0	-0.030	[-0.095, +0.035]	49/50	94.0	2.0	3.000
MP7	12.0	-0.050	[-0.302, +0.202]	50/50	76.0	12.0	4.275
Mild translationese							
FAB	79.5	-0.750	[-0.940, -0.560]	44/44	11.4	9.1	3.216
GEM	76.0	-0.790	[-0.992, -0.588]	50/50	14.0	10.0	3.420
JEV	54.0	+0.006	[-0.060, +0.073]	50/50	0.0	46.0	3.998
GPT	82.0	-0.600	[-0.731, -0.469]	50/50	12.0	6.0	3.560
Q35	52.0	-0.320	[-0.509, -0.131]	50/50	26.0	22.0	3.750
GLM	60.0	-0.360	[-0.521, -0.199]	50/50	26.0	14.0	3.540
DSK	32.0	-0.110	[-0.362, +0.142]	50/50	44.0	24.0	3.950
NMU	36.0	-0.030	[-0.316, +0.256]	49/50	38.0	26.0	2.510
GMA	12.0	+0.070	[-0.023, +0.163]	50/50	66.0	22.0	4.110
Q25	16.0	+0.075	[-0.161, +0.311]	50/50	66.0	18.0	3.925
MP14	4.0	+0.100	[-0.220, +0.420]	49/50	86.0	10.0	1.888
CEJ	12.0	-0.075	[-0.252, +0.102]	50/50	80.0	8.0	4.700
MP3	4.0	+0.475	[+0.080, +0.870]	50/50	76.0	20.0	2.575
NML	2.0	+0.080	[-0.043, +0.203]	49/50	88.0	10.0	3.112
MP7	4.0	+0.200	[-0.014, +0.414]	50/50	82.0	14.0	4.525
Subtle translationese							
FAB	4.5	+0.375	[+0.225, +0.525]	44/44	43.2	52.3	4.341
GEM	2.0	+0.420	[+0.288, +0.552]	50/50	40.0	58.0	4.630
JEV	8.0	+0.232	[+0.165, +0.300]	50/50	0.0	92.0	4.224
GPT	2.0	+0.280	[+0.191, +0.369]	50/50	46.0	52.0	4.440
Q35	10.0	+0.400	[+0.232, +0.568]	50/50	38.0	52.0	4.470
GLM	6.0	+0.470	[+0.344, +0.596]	50/50	22.0	72.0	4.370
DSK	2.0	+0.570	[+0.386, +0.754]	50/50	42.0	56.0	4.630
NMU	24.0	+0.200	[-0.055, +0.455]	49/50	32.0	44.0	2.745
GMA	2.0	+0.240	[+0.150, +0.330]	50/50	54.0	44.0	4.280
Q25	<u>12.0</u>	+0.200	[-0.065, +0.465]	50/50	62.0	26.0	4.050
MP14	2.0	+0.725	[+0.328, +1.122]	49/50	72.0	26.0	2.526
CEJ	2.0	+0.125	[-0.050, +0.300]	50/50	90.0	8.0	4.900
MP3	0.0	+0.775	[+0.380, +1.170]	50/50	74.0	26.0	2.875
NML	2.0	+0.010	[-0.076, +0.096]	49/50	94.0	4.0	3.041
MP7	0.0	+0.300	[+0.062, +0.538]	50/50	86.0	14.0	4.625

A.3.9 Hungarian

Table 13: Hungarian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	87.5	-2.087	$[-2.395, -1.780]$	40/40	10.0	2.5	1.750
GEM	92.0	-1.920	$[-2.189, -1.651]$	50/50	6.0	2.0	2.120
JEV	92.0	-0.859	$[-1.027, -0.691]$	50/50	0.0	8.0	3.011
GPT	92.0	-1.460	$[-1.658, -1.262]$	50/50	4.0	4.0	2.580
Q35	80.0	-1.120	$[-1.391, -0.849]$	50/50	12.0	8.0	2.720
GLM	86.0	-1.120	$[-1.357, -0.883]$	50/50	4.0	10.0	2.530
DSK	76.0	-1.330	$[-1.644, -1.016]$	50/50	18.0	6.0	2.530
NMU	26.0	+0.000	$[-0.260, +0.260]$	45/50	44.0	30.0	1.033
GMA	64.0	-0.370	$[-0.507, -0.233]$	50/50	30.0	6.0	3.840
Q25	16.0	+0.250	$[-0.078, +0.578]$	50/50	54.0	30.0	3.975
MP14	32.0	-0.675	$[-1.166, -0.184]$	50/50	58.0	10.0	1.675
CEJ	6.0	-0.025	$[-0.217, +0.167]$	50/50	86.0	8.0	4.775
MP3	28.0	-0.200	$[-0.716, +0.316]$	50/50	56.0	16.0	2.000
NML	10.0	-0.050	$[-0.159, +0.059]$	50/50	86.0	4.0	2.960
MP7	0.0	+0.350	$[+0.117, +0.583]$	50/50	82.0	18.0	4.825
Mild translationese							
FAB	77.5	-0.700	$[-0.967, -0.433]$	40/40	2.5	20.0	3.138
GEM	68.0	-0.590	$[-0.827, -0.353]$	50/50	18.0	14.0	3.450
JEV	54.0	+0.128	$[-0.002, +0.258]$	50/50	0.0	46.0	3.998
GPT	80.0	-0.530	$[-0.704, -0.356]$	50/50	8.0	12.0	3.510
Q35	38.0	-0.150	$[-0.381, +0.081]$	50/50	22.0	40.0	3.690
GLM	38.0	-0.190	$[-0.390, +0.010]$	50/50	48.0	14.0	3.460
DSK	22.0	+0.210	$[-0.074, +0.494]$	50/50	38.0	40.0	4.070
NMU	24.0	+0.210	$[-0.051, +0.471]$	45/50	30.0	46.0	1.267
GMA	10.0	+0.120	$[+0.003, +0.237]$	50/50	66.0	24.0	4.330
Q25	12.0	+0.325	$[-0.031, +0.681]$	50/50	60.0	28.0	4.050
MP14	14.0	-0.200	$[-0.566, +0.166]$	50/50	74.0	12.0	2.150
CEJ	4.0	+0.100	$[-0.037, +0.237]$	50/50	84.0	12.0	4.900
MP3	16.0	+0.100	$[-0.418, +0.618]$	50/50	68.0	16.0	2.300
NML	4.0	-0.010	$[-0.109, +0.089]$	50/50	92.0	4.0	3.000
MP7	4.0	+0.200	$[-0.002, +0.402]$	50/50	82.0	14.0	4.675
Subtle translationese							
FAB	12.5	+0.400	$[+0.172, +0.628]$	40/40	40.0	47.5	4.237
GEM	4.0	+0.620	$[+0.423, +0.817]$	50/50	32.0	64.0	4.660
JEV	10.0	+0.423	$[+0.254, +0.593]$	50/50	0.0	90.0	4.293
GPT	8.0	+0.380	$[+0.228, +0.532]$	50/50	40.0	52.0	4.420
Q35	4.0	+0.740	$[+0.530, +0.950]$	50/50	28.0	68.0	4.580
GLM	4.0	+0.720	$[+0.512, +0.928]$	50/50	26.0	70.0	4.370
DSK	0.0	+0.910	$[+0.621, +1.199]$	50/50	46.0	54.0	4.770
NMU	24.0	+0.100	$[-0.181, +0.381]$	45/50	42.0	34.0	1.144
GMA	4.0	+0.220	$[+0.085, +0.355]$	50/50	64.0	32.0	4.430
Q25	8.0	+0.450	$[+0.172, +0.728]$	50/50	58.0	34.0	4.175
MP14	4.0	+0.550	$[+0.192, +0.908]$	50/50	68.0	28.0	2.900
CEJ	2.0	+0.025	$[-0.153, +0.203]$	50/50	90.0	8.0	4.825
MP3	6.0	+0.825	$[+0.357, +1.293]$	50/50	68.0	26.0	3.025
NML	12.0	-0.090	$[-0.248, +0.068]$	50/50	82.0	6.0	2.920
MP7	6.0	+0.275	$[+0.020, +0.530]$	50/50	76.0	18.0	4.750

A.3.10 Icelandic

Table 14: Icelandic: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	88.1	-1.929	$[-2.262, -1.595]$	42/42	9.5	2.4	1.750
GEM	86.0	-1.650	$[-1.968, -1.332]$	50/50	6.0	8.0	2.210
JEV	78.0	-0.353	$[-0.501, -0.204]$	50/50	0.0	22.0	3.409
GPT	<u>88.0</u>	-1.280	$[-1.519, -1.041]$	50/50	4.0	8.0	2.640
Q35	78.0	-0.990	$[-1.265, -0.715]$	50/50	10.0	12.0	2.710
GLM	78.0	-1.070	$[-1.306, -0.834]$	50/50	18.0	4.0	2.580
DSK	64.0	-1.100	$[-1.434, -0.766]$	50/50	28.0	8.0	2.580
NMU	20.0	+0.300	$[+0.081, +0.519]$	48/50	40.0	40.0	1.302
GMA	38.0	-0.160	$[-0.307, -0.013]$	50/50	52.0	10.0	3.920
Q25	16.0	+0.200	$[-0.074, +0.474]$	50/50	58.0	26.0	3.750
MP14	18.0	+0.150	$[-0.262, +0.562]$	50/50	62.0	20.0	2.500
CEJ	12.0	+0.025	$[-0.167, +0.217]$	50/50	76.0	12.0	4.625
MP3	18.0	-0.125	$[-0.541, +0.291]$	49/50	68.0	14.0	1.633
NML	0.0	+0.060	$[-0.058, +0.178]$	50/50	98.0	2.0	3.000
MP7	10.0	+0.150	$[-0.089, +0.389]$	50/50	74.0	16.0	4.450
Mild translationese							
FAB	66.7	-0.690	$[-0.983, -0.397]$	42/42	7.1	26.2	2.988
GEM	62.0	-0.410	$[-0.687, -0.133]$	50/50	14.0	24.0	3.450
JEV	26.0	+0.236	$[+0.123, +0.348]$	50/50	2.0	72.0	3.997
GPT	<u>66.0</u>	-0.420	$[-0.619, -0.221]$	50/50	14.0	20.0	3.500
Q35	34.0	+0.150	$[-0.136, +0.436]$	50/50	24.0	42.0	3.850
GLM	36.0	+0.090	$[-0.105, +0.285]$	50/50	30.0	34.0	3.740
DSK	24.0	+0.350	$[-0.014, +0.714]$	50/50	38.0	38.0	4.030
NMU	28.0	+0.330	$[+0.033, +0.627]$	48/50	24.0	48.0	1.333
GMA	4.0	+0.190	$[+0.072, +0.308]$	50/50	66.0	30.0	4.270
Q25	14.0	+0.225	$[-0.108, +0.558]$	50/50	60.0	26.0	3.775
MP14	8.0	+0.625	$[+0.234, +1.016]$	50/50	58.0	34.0	2.975
CEJ	8.0	+0.100	$[-0.083, +0.283]$	50/50	78.0	14.0	4.700
MP3	6.0	+0.350	$[-0.000, +0.700]$	49/50	76.0	18.0	2.117
NML	0.0	+0.130	$[-0.020, +0.280]$	50/50	94.0	6.0	3.070
MP7	10.0	+0.100	$[-0.120, +0.320]$	50/50	76.0	14.0	4.400
Subtle translationese							
FAB	9.5	+0.631	$[+0.355, +0.906]$	42/42	38.1	52.4	4.310
GEM	6.0	+0.810	$[+0.526, +1.094]$	50/50	26.0	68.0	4.670
JEV	<u>14.0</u>	+0.350	$[+0.237, +0.463]$	50/50	0.0	86.0	4.112
GPT	10.0	+0.480	$[+0.286, +0.674]$	50/50	38.0	52.0	4.400
Q35	6.0	+0.850	$[+0.622, +1.078]$	50/50	22.0	72.0	4.550
GLM	10.0	+0.620	$[+0.388, +0.852]$	50/50	24.0	66.0	4.270
DSK	2.0	+0.940	$[+0.640, +1.240]$	50/50	34.0	64.0	4.620
NMU	24.0	+0.370	$[+0.109, +0.631]$	48/50	24.0	52.0	1.375
GMA	0.0	+0.300	$[+0.178, +0.422]$	50/50	58.0	42.0	4.380
Q25	<u>14.0</u>	+0.225	$[-0.156, +0.606]$	50/50	60.0	26.0	3.775
MP14	8.0	+0.450	$[+0.075, +0.825]$	50/50	62.0	30.0	2.800
CEJ	<u>14.0</u>	+0.050	$[-0.160, +0.260]$	50/50	70.0	16.0	4.650
MP3	8.0	+0.375	$[-0.004, +0.754]$	49/50	74.0	18.0	2.143
NML	0.0	+0.100	$[-0.031, +0.231]$	50/50	94.0	6.0	3.040
MP7	10.0	<u>+0.075</u>	$[-0.161, +0.311]$	50/50	82.0	8.0	4.375

A.3.11 Irish

Table 15: Irish: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	42.5	-0.487	[-0.918, -0.057]	40/40	25.0	32.5	2.350
GEM	34.0	-0.200	[-0.653, +0.253]	50/50	26.0	40.0	2.660
JEV	34.0	+0.208	[+0.002, +0.414]	50/50	0.0	66.0	3.383
GPT	24.0	+0.190	[-0.032, +0.412]	50/50	32.0	44.0	3.400
Q35	24.0	+0.170	[-0.141, +0.481]	50/50	34.0	42.0	3.070
GLM	22.0	+0.250	[+0.054, +0.446]	50/50	36.0	42.0	3.330
DSK	24.0	+0.360	[+0.000, +0.720]	50/50	30.0	46.0	3.370
NMU	16.0	-0.010	[-0.134, +0.114]	48/50	76.0	8.0	0.104
GMA	16.0	+0.100	[-0.136, +0.336]	50/50	36.0	48.0	3.520
Q25	30.0	-0.125	[-0.516, +0.266]	49/50	40.0	30.0	3.316
MP14	28.0	-0.150	[-0.562, +0.262]	50/50	52.0	20.0	1.800
CEJ	32.0	-0.375	[-0.796, +0.046]	50/50	46.0	22.0	3.925
MP3	10.0	+0.550	[-0.081, +1.181]	49/50	62.0	28.0	2.985
NML	28.0	-0.690	[-1.068, -0.312]	50/50	64.0	8.0	2.020
MP7	18.0	+0.150	[-0.296, +0.596]	50/50	56.0	26.0	4.100
Mild translationese							
FAB	45.0	-0.613	[-1.136, -0.089]	40/40	25.0	30.0	2.225
GEM	30.0	+0.060	[-0.454, +0.574]	50/50	18.0	52.0	2.920
JEV	24.0	+0.431	[+0.235, +0.627]	50/50	0.0	76.0	3.606
GPT	22.0	+0.370	[+0.126, +0.614]	50/50	24.0	54.0	3.580
Q35	18.0	+0.450	[+0.130, +0.770]	50/50	22.0	60.0	3.350
GLM	14.0	+0.550	[+0.329, +0.771]	50/50	28.0	58.0	3.630
DSK	12.0	+0.690	[+0.307, +1.073]	50/50	34.0	54.0	3.700
NMU	16.0	+0.040	[-0.110, +0.190]	48/50	72.0	12.0	0.156
GMA	12.0	+0.280	[+0.086, +0.474]	50/50	38.0	50.0	3.700
Q25	40.0	-0.400	[-0.792, -0.008]	49/50	38.0	22.0	3.036
MP14	26.0	-0.050	[-0.498, +0.398]	50/50	52.0	22.0	1.900
CEJ	36.0	-0.600	[-1.101, -0.099]	50/50	38.0	26.0	3.700
MP3	8.0	+0.600	[-0.065, +1.265]	49/50	66.0	26.0	3.036
NML	26.0	-0.510	[-0.885, -0.135]	50/50	66.0	8.0	2.200
MP7	22.0	+0.125	[-0.302, +0.552]	50/50	54.0	24.0	4.075
Subtle translationese							
FAB	22.5	+0.163	[-0.370, +0.695]	40/40	17.5	60.0	3.000
GEM	20.0	+0.370	[-0.063, +0.803]	50/50	22.0	58.0	3.230
JEV	24.0	+0.600	[+0.396, +0.804]	50/50	0.0	76.0	3.775
GPT	4.0	+0.710	[+0.479, +0.941]	50/50	30.0	66.0	3.920
Q35	12.0	+0.670	[+0.318, +1.022]	50/50	16.0	72.0	3.570
GLM	2.0	+0.850	[+0.622, +1.078]	50/50	26.0	72.0	3.930
DSK	8.0	+1.080	[+0.751, +1.409]	50/50	18.0	74.0	4.090
NMU	14.0	+0.040	[-0.078, +0.158]	48/50	70.0	16.0	0.156
GMA	12.0	+0.340	[+0.081, +0.599]	50/50	26.0	62.0	3.760
Q25	30.0	-0.325	[-0.783, +0.133]	49/50	46.0	24.0	3.112
MP14	28.0	+0.025	[-0.515, +0.565]	50/50	44.0	28.0	1.975
CEJ	30.0	-0.450	[-0.973, +0.073]	50/50	44.0	26.0	3.850
MP3	12.0	+0.475	[-0.174, +1.124]	49/50	62.0	26.0	2.908
NML	24.0	-0.540	[-0.920, -0.160]	50/50	66.0	10.0	2.170
MP7	16.0	+0.325	[-0.094, +0.744]	50/50	54.0	30.0	4.275

A.3.12 Italian

Table 16: Italian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	<u>97.4</u>	-2.218	[-2.453, -1.982]	39/39	0.0	2.6	1.885
GEM	98.0	-1.920	[-2.117, -1.723]	50/50	2.0	0.0	2.250
JEV	92.0	-1.002	[-1.173, -0.831]	50/50	0.0	8.0	3.033
GPT	98.0	-1.390	[-1.536, -1.244]	50/50	2.0	0.0	2.720
Q35	88.0	-1.280	[-1.515, -1.045]	50/50	6.0	6.0	2.770
GLM	92.0	-1.460	[-1.660, -1.260]	50/50	8.0	0.0	2.500
DSK	88.0	-1.660	[-1.913, -1.407]	50/50	12.0	0.0	2.490
NMU	48.0	-0.460	[-0.680, -0.240]	45/50	38.0	14.0	2.656
GMA	82.0	-0.620	[-0.753, -0.487]	50/50	14.0	4.0	3.570
Q25	18.0	-0.150	[-0.356, +0.056]	50/50	76.0	6.0	3.175
MP14	50.0	-1.325	[-1.764, -0.886]	50/50	48.0	2.0	1.200
CEJ	22.0	-0.225	[-0.419, -0.031]	50/50	72.0	6.0	4.525
MP3	46.0	-0.975	[-1.441, -0.509]	50/50	50.0	4.0	0.750
NML	12.0	-0.080	[-0.170, +0.010]	50/50	84.0	4.0	3.000
MP7	26.0	-0.450	[-0.784, -0.116]	50/50	72.0	2.0	3.975
Mild translationese							
FAB	87.2	-1.192	[-1.422, -0.963]	39/39	10.3	2.6	2.910
GEM	<u>86.0</u>	-0.940	[-1.153, -0.727]	50/50	6.0	8.0	3.230
JEV	70.0	-0.049	[-0.179, +0.080]	50/50	0.0	30.0	3.986
GPT	84.0	-0.740	[-0.883, -0.597]	50/50	12.0	4.0	3.370
Q35	58.0	-0.460	[-0.685, -0.235]	50/50	22.0	20.0	3.590
GLM	72.0	-0.620	[-0.840, -0.400]	50/50	10.0	18.0	3.340
DSK	50.0	-0.470	[-0.752, -0.188]	50/50	36.0	14.0	3.680
NMU	30.0	-0.140	[-0.340, +0.060]	45/50	44.0	26.0	3.011
GMA	18.0	+0.040	[-0.071, +0.151]	50/50	60.0	22.0	4.230
Q25	12.0	-0.025	[-0.203, +0.153]	50/50	80.0	8.0	3.300
MP14	14.0	-0.225	[-0.587, +0.137]	50/50	76.0	10.0	2.300
CEJ	8.0	-0.050	[-0.271, +0.171]	50/50	84.0	8.0	4.700
MP3	18.0	-0.275	[-0.647, +0.097]	50/50	76.0	6.0	1.450
NML	8.0	+0.040	[-0.068, +0.148]	50/50	82.0	10.0	3.120
MP7	10.0	+0.050	[-0.147, +0.247]	50/50	80.0	10.0	4.475
Subtle translationese							
FAB	5.1	+0.385	[+0.158, +0.611]	39/39	46.2	48.7	4.487
GEM	4.0	+0.630	[+0.432, +0.828]	50/50	24.0	72.0	4.800
JEV	10.0	+0.335	[+0.207, +0.463]	50/50	0.0	90.0	4.370
GPT	4.0	+0.370	[+0.245, +0.495]	50/50	36.0	60.0	4.480
Q35	<u>12.0</u>	+0.560	[+0.365, +0.755]	50/50	22.0	66.0	4.610
GLM	4.0	+0.570	[+0.380, +0.760]	50/50	32.0	64.0	4.530
DSK	2.0	+0.700	[+0.448, +0.952]	50/50	46.0	52.0	4.850
NMU	16.0	+0.310	[+0.086, +0.534]	45/50	42.0	42.0	3.511
GMA	2.0	+0.270	[+0.141, +0.399]	50/50	62.0	36.0	4.460
Q25	10.0	<u>+0.000</u>	[-0.198, +0.198]	50/50	80.0	10.0	3.325
MP14	2.0	+0.525	[+0.181, +0.869]	50/50	72.0	26.0	3.050
CEJ	<u>12.0</u>	-0.050	[-0.235, +0.135]	50/50	78.0	10.0	4.700
MP3	2.0	+0.750	[+0.342, +1.158]	50/50	68.0	30.0	2.475
NML	6.0	+0.450	[+0.231, +0.669]	50/50	62.0	32.0	3.530
MP7	10.0	+0.200	[-0.083, +0.483]	50/50	76.0	14.0	4.625

A.3.13 Latvian

Table 17: Latvian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	87.2	-1.487	[-1.779, -1.196]	39/39	10.3	2.6	1.769
GEM	88.0	-1.350	[-1.626, -1.074]	49/50	10.0	2.0	2.031
JEV	88.0	-0.556	[-0.703, -0.409]	50/50	0.0	12.0	3.115
GPT	92.0	-1.020	[-1.197, -0.843]	50/50	6.0	2.0	2.570
Q35	64.0	-0.760	[-1.004, -0.516]	50/50	26.0	10.0	2.770
GLM	80.0	-0.860	[-1.068, -0.652]	50/50	16.0	4.0	2.540
DSK	66.0	-1.090	[-1.389, -0.791]	50/50	28.0	6.0	2.450
NMU	18.0	+0.250	[+0.027, +0.473]	50/50	42.0	40.0	1.070
GMA	74.0	-0.410	[-0.506, -0.314]	50/50	22.0	4.0	3.620
Q25	20.0	+0.225	[-0.143, +0.593]	49/50	54.0	26.0	3.929
MP14	20.0	-0.350	[-0.663, -0.037]	50/50	72.0	8.0	1.500
CEJ	18.0	-0.075	[-0.331, +0.181]	50/50	70.0	12.0	4.650
MP3	16.0	-0.025	[-0.465, +0.415]	50/50	74.0	10.0	3.050
NML	6.0	-0.020	[-0.195, +0.155]	50/50	90.0	4.0	2.910
MP7	20.0	-0.125	[-0.510, +0.260]	50/50	70.0	10.0	4.000
Mild translationese							
FAB	53.8	-0.359	[-0.626, -0.092]	39/39	33.3	12.8	2.897
GEM	54.0	-0.330	[-0.578, -0.082]	49/50	22.0	24.0	3.071
JEV	44.0	+0.117	[-0.000, +0.234]	50/50	2.0	54.0	3.788
GPT	50.0	-0.290	[-0.461, -0.119]	50/50	34.0	16.0	3.300
Q35	36.0	-0.090	[-0.301, +0.121]	50/50	28.0	36.0	3.440
GLM	50.0	-0.140	[-0.346, +0.066]	50/50	20.0	30.0	3.260
DSK	38.0	-0.030	[-0.312, +0.252]	50/50	34.0	28.0	3.510
NMU	42.0	+0.040	[-0.237, +0.317]	50/50	16.0	42.0	0.860
GMA	10.0	+0.060	[-0.027, +0.147]	50/50	72.0	18.0	4.090
Q25	8.0	+0.225	[-0.086, +0.536]	49/50	70.0	22.0	3.929
MP14	8.0	-0.075	[-0.350, +0.200]	50/50	86.0	6.0	1.775
CEJ	18.0	-0.025	[-0.272, +0.222]	50/50	72.0	10.0	4.700
MP3	8.0	+0.100	[-0.313, +0.513]	50/50	78.0	14.0	3.175
NML	8.0	-0.050	[-0.216, +0.116]	50/50	90.0	2.0	2.880
MP7	14.0	-0.025	[-0.342, +0.292]	50/50	78.0	8.0	4.100
Subtle translationese							
FAB	5.1	+0.744	[+0.487, +1.000]	39/39	25.6	69.2	4.000
GEM	4.0	+0.710	[+0.488, +0.932]	49/50	32.0	64.0	4.133
JEV	<u>14.0</u>	+0.370	[+0.235, +0.506]	50/50	2.0	84.0	4.042
GPT	2.0	+0.520	[+0.369, +0.671]	50/50	30.0	68.0	4.110
Q35	8.0	+0.600	[+0.394, +0.806]	50/50	32.0	60.0	4.130
GLM	6.0	+0.620	[+0.417, +0.823]	50/50	24.0	70.0	4.020
DSK	10.0	+0.690	[+0.416, +0.964]	50/50	38.0	52.0	4.230
NMU	30.0	+0.290	[-0.022, +0.602]	50/50	32.0	38.0	1.110
GMA	2.0	+0.230	[+0.121, +0.339]	50/50	64.0	34.0	4.260
Q25	<u>14.0</u>	+0.300	[+0.007, +0.593]	49/50	58.0	28.0	4.005
MP14	0.0	+0.375	[+0.130, +0.620]	50/50	80.0	20.0	2.225
CEJ	10.0	+0.025	[-0.153, +0.203]	50/50	80.0	10.0	4.750
MP3	2.0	+0.525	[+0.116, +0.934]	50/50	78.0	20.0	3.600
NML	4.0	+0.000	[-0.170, +0.170]	50/50	92.0	4.0	2.930
MP7	<u>14.0</u>	+0.150	[-0.160, +0.460]	50/50	74.0	12.0	4.275

A.3.14 Limburgish

Table 18: Limburgish: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	75.0	-1.528	$[-2.084, -0.972]$	36/36	19.4	5.6	2.069
GEM	<u>68.0</u>	<u>-1.390</u>	$[-1.900, -0.880]$	50/50	26.0	6.0	2.820
JEV	52.0	+0.018	$[-0.099, +0.136]$	50/50	2.0	46.0	3.688
GPT	60.0	-0.450	$[-0.728, -0.172]$	50/50	20.0	20.0	3.460
Q35	60.0	-0.970	$[-1.458, -0.482]$	50/50	18.0	22.0	2.720
GLM	42.0	-0.150	$[-0.489, +0.189]$	50/50	34.0	24.0	3.690
DSK	50.0	-0.920	$[-1.391, -0.449]$	50/50	30.0	20.0	3.410
NMU	32.0	+0.090	$[-0.375, +0.555]$	45/50	22.0	46.0	1.722
GMA	20.0	-0.180	$[-0.376, +0.016]$	50/50	72.0	8.0	4.050
Q25	32.0	-0.300	$[-0.706, +0.106]$	50/50	48.0	20.0	2.825
MP14	34.0	-0.650	$[-1.249, -0.051]$	50/50	50.0	16.0	2.575
CEJ	24.0	+0.025	$[-0.404, +0.454]$	50/50	54.0	22.0	4.425
MP3	12.0	+0.700	$[+0.093, +1.307]$	49/50	62.0	26.0	2.143
NML	30.0	-0.310	$[-0.846, +0.226]$	50/50	50.0	20.0	2.240
MP7	18.0	-0.125	$[-0.360, +0.110]$	50/50	76.0	6.0	4.600
Mild translationese							
FAB	<u>63.9</u>	-1.361	$[-1.907, -0.815]$	36/36	27.8	8.3	2.236
GEM	74.0	-1.260	$[-1.707, -0.813]$	50/50	22.0	4.0	2.950
JEV	36.0	+0.114	$[+0.010, +0.217]$	50/50	4.0	60.0	3.783
GPT	56.0	-0.330	$[-0.588, -0.072]$	50/50	28.0	16.0	3.580
Q35	38.0	-0.290	$[-0.740, +0.160]$	50/50	28.0	34.0	3.400
GLM	40.0	+0.000	$[-0.257, +0.257]$	50/50	32.0	28.0	3.840
DSK	52.0	-0.880	$[-1.345, -0.415]$	50/50	34.0	14.0	3.450
NMU	32.0	-0.030	$[-0.352, +0.292]$	45/50	34.0	34.0	1.589
GMA	14.0	-0.030	$[-0.163, +0.103]$	50/50	70.0	16.0	4.200
Q25	26.0	-0.200	$[-0.650, +0.250]$	50/50	48.0	26.0	2.925
MP14	26.0	-0.200	$[-0.650, +0.250]$	50/50	56.0	18.0	3.025
CEJ	26.0	+0.050	$[-0.358, +0.458]$	50/50	48.0	26.0	4.450
MP3	16.0	+0.725	$[+0.152, +1.298]$	49/50	54.0	30.0	2.168
NML	24.0	-0.360	$[-0.796, +0.076]$	50/50	58.0	18.0	2.190
MP7	14.0	-0.075	$[-0.301, +0.151]$	50/50	80.0	6.0	4.650
Subtle translationese							
FAB	52.8	-1.042	$[-1.582, -0.502]$	36/36	38.9	8.3	2.556
GEM	36.0	-0.640	$[-1.109, -0.171]$	50/50	54.0	10.0	3.570
JEV	16.0	+0.173	$[+0.066, +0.280]$	50/50	8.0	76.0	3.843
GPT	20.0	+0.060	$[-0.133, +0.253]$	50/50	60.0	20.0	3.970
Q35	34.0	-0.100	$[-0.521, +0.321]$	50/50	28.0	38.0	3.590
GLM	22.0	+0.180	$[-0.044, +0.404]$	50/50	38.0	40.0	4.020
DSK	30.0	-0.500	$[-0.947, -0.053]$	50/50	46.0	24.0	3.830
NMU	<u>38.0</u>	+0.140	$[-0.310, +0.590]$	45/50	22.0	40.0	1.778
GMA	10.0	+0.050	$[-0.044, +0.144]$	50/50	72.0	18.0	4.280
Q25	18.0	+0.050	$[-0.293, +0.393]$	50/50	58.0	24.0	3.175
MP14	24.0	-0.050	$[-0.559, +0.459]$	50/50	56.0	20.0	3.175
CEJ	18.0	+0.000	$[-0.364, +0.364]$	50/50	60.0	22.0	4.400
MP3	12.0	+0.525	$[+0.001, +1.049]$	49/50	70.0	18.0	1.964
NML	26.0	-0.350	$[-0.793, +0.093]$	50/50	56.0	18.0	2.200
MP7	10.0	-0.050	$[-0.235, +0.135]$	50/50	88.0	2.0	4.675

A.3.15 Lower Sorbian

Table 19: Lower Sorbian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	75.6	-1.024	[-1.420, -0.628]	41/41	14.6	9.8	2.537
GEM	48.0	-0.460	[-0.751, -0.169]	50/50	26.0	26.0	3.440
JEV	76.0	-0.329	[-0.486, -0.173]	50/50	0.0	24.0	3.084
GPT	58.0	-0.350	[-0.548, -0.152]	50/50	26.0	16.0	3.330
Q35	48.0	-0.950	[-1.438, -0.462]	50/50	36.0	16.0	2.530
GLM	56.0	-0.310	[-0.506, -0.114]	50/50	22.0	22.0	3.360
DSK	62.0	-1.130	[-1.617, -0.643]	50/50	22.0	16.0	2.290
NMU	44.0	-0.310	[-0.768, +0.148]	48/50	20.0	36.0	1.438
GMA	34.0	-0.680	[-1.040, -0.320]	50/50	50.0	16.0	3.010
Q25	14.0	+0.550	[+0.118, +0.982]	50/50	50.0	36.0	3.200
MP14	24.0	-0.250	[-0.633, +0.133]	50/50	66.0	10.0	1.450
CEJ	18.0	+0.050	[-0.307, +0.407]	50/50	64.0	18.0	4.425
MP3	12.0	+1.200	[+0.602, +1.798]	50/50	50.0	38.0	1.975
NML	34.0	-0.220	[-0.699, +0.259]	50/50	42.0	24.0	1.240
MP7	20.0	+0.000	[-0.221, +0.221]	50/50	60.0	20.0	4.425
Mild translationese							
FAB	63.4	-0.878	[-1.275, -0.481]	41/41	24.4	12.2	2.683
GEM	50.0	-0.220	[-0.523, +0.083]	50/50	12.0	38.0	3.680
JEV	64.0	-0.307	[-0.464, -0.150]	50/50	4.0	32.0	3.106
GPT	52.0	-0.190	[-0.392, +0.012]	50/50	28.0	20.0	3.490
Q35	48.0	-0.510	[-0.946, -0.074]	50/50	24.0	28.0	2.970
GLM	44.0	-0.020	[-0.242, +0.202]	50/50	24.0	32.0	3.650
DSK	60.0	-0.900	[-1.399, -0.401]	50/50	16.0	24.0	2.520
NMU	54.0	-0.310	[-0.794, +0.174]	48/50	12.0	34.0	1.438
GMA	30.0	-0.680	[-1.066, -0.294]	50/50	50.0	20.0	3.010
Q25	12.0	+0.625	[+0.204, +1.046]	50/50	52.0	36.0	3.275
MP14	30.0	-0.350	[-0.758, +0.058]	50/50	54.0	16.0	1.350
CEJ	20.0	+0.200	[-0.186, +0.586]	50/50	54.0	26.0	4.575
MP3	6.0	+1.325	[+0.750, +1.900]	50/50	52.0	42.0	2.100
NML	28.0	-0.030	[-0.508, +0.448]	50/50	38.0	34.0	1.430
MP7	14.0	+0.125	[-0.088, +0.338]	50/50	62.0	24.0	4.550
Subtle translationese							
FAB	34.1	-0.585	[-1.080, -0.091]	41/41	41.5	24.4	2.976
GEM	26.0	+0.080	[-0.198, +0.358]	50/50	26.0	48.0	3.980
JEV	48.0	-0.193	[-0.362, -0.024]	50/50	4.0	48.0	3.220
GPT	28.0	+0.190	[-0.002, +0.382]	50/50	32.0	40.0	3.870
Q35	32.0	-0.370	[-0.836, +0.096]	50/50	32.0	36.0	3.110
GLM	28.0	+0.230	[+0.018, +0.442]	50/50	22.0	50.0	3.900
DSK	56.0	-0.690	[-1.242, -0.138]	50/50	16.0	28.0	2.730
NMU	46.0	-0.070	[-0.597, +0.457]	48/50	18.0	36.0	1.688
GMA	30.0	-0.640	[-1.014, -0.266]	50/50	52.0	18.0	3.050
Q25	12.0	+0.600	[+0.162, +1.038]	50/50	56.0	32.0	3.250
MP14	20.0	-0.050	[-0.458, +0.358]	50/50	64.0	16.0	1.650
CEJ	8.0	+0.225	[-0.024, +0.474]	50/50	74.0	18.0	4.600
MP3	8.0	+1.200	[+0.623, +1.777]	50/50	56.0	36.0	1.975
NML	30.0	+0.060	[-0.424, +0.544]	50/50	42.0	28.0	1.520
MP7	12.0	+0.150	[-0.067, +0.367]	50/50	66.0	22.0	4.575

A.3.16 Luxembourgish

Table 20: Luxembourgish: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	73.8	-1.524	$[-2.060, -0.987]$	42/42	14.3	11.9	2.202
GEM	64.0	-1.330	$[-1.834, -0.826]$	50/50	20.0	16.0	2.750
JEV	58.0	-0.103	$[-0.243, +0.037]$	50/50	0.0	42.0	3.198
GPT	50.0	-0.200	$[-0.381, -0.019]$	50/50	28.0	22.0	3.610
Q35	42.0	-0.670	$[-1.125, -0.215]$	49/50	28.0	30.0	2.663
GLM	34.0	+0.110	$[-0.096, +0.316]$	50/50	28.0	38.0	3.530
DSK	42.0	-0.700	$[-1.224, -0.176]$	50/50	30.0	28.0	2.940
NMU	50.0	-0.440	$[-0.883, +0.003]$	47/50	22.0	28.0	1.457
GMA	34.0	-0.680	$[-1.056, -0.304]$	50/50	48.0	18.0	3.410
Q25	30.0	-0.300	$[-0.706, +0.106]$	48/50	44.0	26.0	3.255
MP14	36.0	-0.750	$[-1.288, -0.212]$	50/50	54.0	10.0	1.975
CEJ	28.0	-0.175	$[-0.472, +0.122]$	50/50	58.0	14.0	4.550
MP3	20.0	+0.125	$[-0.466, +0.716]$	50/50	58.0	22.0	2.300
NML	28.0	+0.010	$[-0.478, +0.498]$	50/50	42.0	30.0	2.270
MP7	18.0	-0.375	$[-0.766, +0.016]$	50/50	72.0	10.0	4.425
Mild translationese							
FAB	69.0	-1.238	$[-1.805, -0.671]$	42/42	9.5	21.4	2.488
GEM	64.0	-1.010	$[-1.475, -0.545]$	50/50	18.0	18.0	3.070
JEV	44.0	+0.010	$[-0.123, +0.143]$	50/50	0.0	56.0	3.311
GPT	42.0	-0.070	$[-0.238, +0.098]$	50/50	28.0	30.0	3.740
Q35	48.0	-0.780	$[-1.281, -0.279]$	49/50	22.0	30.0	2.551
GLM	12.0	+0.510	$[+0.307, +0.713]$	50/50	30.0	58.0	3.930
DSK	36.0	-0.410	$[-0.931, +0.111]$	50/50	30.0	34.0	3.230
NMU	52.0	-0.270	$[-0.731, +0.191]$	47/50	12.0	36.0	1.638
GMA	34.0	-0.600	$[-0.973, -0.227]$	50/50	42.0	24.0	3.490
Q25	34.0	-0.100	$[-0.463, +0.263]$	48/50	32.0	34.0	3.464
MP14	26.0	-0.500	$[-1.014, +0.014]$	50/50	60.0	14.0	2.225
CEJ	20.0	-0.100	$[-0.387, +0.187]$	50/50	70.0	10.0	4.625
MP3	20.0	+0.325	$[-0.323, +0.973]$	50/50	50.0	30.0	2.500
NML	14.0	+0.100	$[-0.318, +0.518]$	50/50	56.0	30.0	2.360
MP7	16.0	-0.450	$[-0.851, -0.049]$	50/50	78.0	6.0	4.350
Subtle translationese							
FAB	50.0	-0.940	$[-1.587, -0.294]$	42/42	19.0	31.0	2.786
GEM	30.0	-0.510	$[-1.059, +0.039]$	50/50	30.0	40.0	3.570
JEV	32.0	+0.152	$[-0.008, +0.311]$	50/50	0.0	68.0	3.453
GPT	12.0	+0.350	$[+0.207, +0.493]$	50/50	32.0	56.0	4.160
Q35	30.0	-0.220	$[-0.788, +0.348]$	49/50	18.0	52.0	3.122
GLM	4.0	+0.720	$[+0.528, +0.912]$	50/50	24.0	72.0	4.140
DSK	30.0	-0.150	$[-0.730, +0.430]$	50/50	28.0	42.0	3.490
NMU	50.0	-0.290	$[-0.756, +0.176]$	47/50	22.0	28.0	1.617
GMA	30.0	-0.560	$[-0.962, -0.158]$	50/50	44.0	26.0	3.530
Q25	36.0	-0.250	$[-0.627, +0.127]$	48/50	36.0	28.0	3.307
MP14	34.0	-0.325	$[-0.981, +0.331]$	50/50	42.0	24.0	2.400
CEJ	28.0	-0.400	$[-0.780, -0.020]$	50/50	62.0	10.0	4.325
MP3	20.0	+0.250	$[-0.348, +0.848]$	50/50	54.0	26.0	2.425
NML	22.0	+0.070	$[-0.380, +0.520]$	50/50	48.0	30.0	2.330
MP7	16.0	-0.400	$[-0.792, -0.008]$	50/50	78.0	6.0	4.400

A.3.17 Maltese

Table 21: Maltese: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	$\Delta \downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	52.5	-0.500	$[-0.824, -0.176]$	40/40	25.0	22.5	3.275
GEM	52.0	-0.700	$[-1.067, -0.333]$	50/50	36.0	12.0	3.290
JEV	44.0	-0.071	$[-0.251, +0.109]$	50/50	2.0	54.0	3.583
GPT	52.0	-0.340	$[-0.516, -0.164]$	50/50	32.0	16.0	3.610
Q35	34.0	-0.300	$[-0.724, +0.124]$	49/50	30.0	36.0	3.510
GLM	36.0	-0.120	$[-0.329, +0.089]$	50/50	36.0	28.0	3.530
DSK	34.0	-0.150	$[-0.565, +0.265]$	50/50	28.0	38.0	3.630
NMU	34.0	+0.010	$[-0.365, +0.385]$	49/50	20.0	46.0	1.480
GMA	26.0	-0.150	$[-0.307, +0.007]$	50/50	62.0	12.0	4.010
Q25	26.0	-0.150	$[-0.452, +0.152]$	46/50	54.0	20.0	4.239
MP14	28.0	-0.550	$[-1.026, -0.074]$	50/50	56.0	16.0	3.200
CEJ	16.0	-0.050	$[-0.420, +0.320]$	50/50	70.0	14.0	4.650
MP3	18.0	-0.125	$[-0.744, +0.494]$	50/50	66.0	16.0	3.400
NML	4.0	-0.060	$[-0.265, +0.145]$	50/50	94.0	2.0	2.880
MP7	14.0	-0.300	$[-0.593, -0.007]$	50/50	84.0	2.0	4.650
Mild translationese							
FAB	37.5	-0.225	$[-0.608, +0.158]$	40/40	30.0	32.5	3.550
GEM	36.0	-0.340	$[-0.692, +0.012]$	50/50	32.0	32.0	3.650
JEV	38.0	+0.022	$[-0.139, +0.183]$	50/50	0.0	62.0	3.676
GPT	34.0	-0.200	$[-0.429, +0.029]$	50/50	44.0	22.0	3.750
Q35	32.0	-0.090	$[-0.461, +0.281]$	49/50	30.0	38.0	3.724
GLM	40.0	-0.280	$[-0.662, +0.102]$	50/50	24.0	36.0	3.370
DSK	18.0	+0.220	$[-0.074, +0.514]$	50/50	44.0	38.0	4.000
NMU	42.0	-0.030	$[-0.403, +0.343]$	49/50	22.0	36.0	1.439
GMA	22.0	-0.180	$[-0.374, +0.014]$	50/50	62.0	16.0	3.980
Q25	22.0	-0.125	$[-0.416, +0.166]$	46/50	60.0	18.0	4.266
MP14	28.0	-0.525	$[-0.995, -0.055]$	50/50	56.0	16.0	3.225
CEJ	16.0	-0.100	$[-0.420, +0.220]$	50/50	68.0	16.0	4.600
MP3	18.0	-0.025	$[-0.633, +0.583]$	50/50	66.0	16.0	3.500
NML	14.0	-0.280	$[-0.567, +0.007]$	50/50	84.0	2.0	2.660
MP7	8.0	-0.200	$[-0.492, +0.092]$	50/50	88.0	4.0	4.750
Subtle translationese							
FAB	22.5	+0.013	$[-0.381, +0.406]$	40/40	35.0	42.5	3.788
GEM	24.0	-0.190	$[-0.695, +0.315]$	50/50	20.0	56.0	3.800
JEV	26.0	+0.100	$[-0.092, +0.292]$	50/50	0.0	74.0	3.754
GPT	18.0	+0.170	$[+0.001, +0.339]$	50/50	38.0	44.0	4.120
Q35	12.0	+0.330	$[-0.048, +0.708]$	49/50	28.0	60.0	4.153
GLM	18.0	+0.260	$[-0.057, +0.577]$	50/50	26.0	56.0	3.910
DSK	16.0	+0.540	$[+0.173, +0.907]$	50/50	28.0	56.0	4.320
NMU	40.0	-0.030	$[-0.398, +0.338]$	49/50	16.0	44.0	1.439
GMA	14.0	-0.090	$[-0.313, +0.133]$	50/50	58.0	28.0	4.070
Q25	24.0	-0.100	$[-0.387, +0.187]$	46/50	54.0	22.0	4.293
MP14	16.0	-0.250	$[-0.682, +0.182]$	50/50	72.0	12.0	3.500
CEJ	8.0	-0.075	$[-0.428, +0.278]$	50/50	82.0	10.0	4.625
MP3	16.0	+0.350	$[-0.223, +0.923]$	50/50	62.0	22.0	3.875
NML	14.0	-0.290	$[-0.597, +0.017]$	50/50	82.0	4.0	2.650
MP7	10.0	-0.175	$[-0.396, +0.046]$	50/50	88.0	2.0	4.775

A.3.18 Portuguese

Table 22: Portuguese: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	89.3	-1.786	$[-2.164, -1.407]$	28/28	3.6	7.1	2.036
GEM	<u>86.0</u>	<u>-1.400</u>	$[-1.673, -1.127]$	50/50	10.0	4.0	2.390
JEV	80.0	-0.647	$[-0.844, -0.449]$	50/50	0.0	20.0	2.901
GPT	<u>86.0</u>	-0.920	$[-1.113, -0.727]$	50/50	8.0	6.0	2.780
Q35	72.0	-0.810	$[-1.052, -0.568]$	50/50	18.0	10.0	2.820
GLM	78.0	-0.820	$[-1.038, -0.602]$	50/50	10.0	12.0	2.570
DSK	70.0	-1.020	$[-1.314, -0.726]$	50/50	20.0	10.0	2.540
NMU	50.0	-0.460	$[-0.699, -0.221]$	47/50	36.0	14.0	2.436
GMA	62.0	-0.380	$[-0.516, -0.244]$	50/50	26.0	12.0	3.480
Q25	8.0	+0.175	$[-0.023, +0.373]$	49/50	72.0	20.0	3.316
MP14	20.0	-0.400	$[-0.724, -0.076]$	50/50	76.0	4.0	1.200
CEJ	36.0	-0.325	$[-0.687, +0.037]$	49/50	52.0	12.0	4.005
MP3	32.0	-0.575	$[-0.953, -0.197]$	50/50	60.0	8.0	0.575
NML	22.0	-0.190	$[-0.394, +0.014]$	50/50	72.0	6.0	2.920
MP7	18.0	-0.175	$[-0.446, +0.096]$	50/50	76.0	6.0	2.825
Mild translationese							
FAB	75.0	-0.589	$[-0.949, -0.229]$	28/28	7.1	17.9	3.232
GEM	<u>54.0</u>	<u>-0.350</u>	$[-0.590, -0.110]$	50/50	22.0	24.0	3.440
JEV	44.0	+0.162	$[+0.008, +0.316]$	50/50	2.0	54.0	3.709
GPT	52.0	-0.230	$[-0.410, -0.050]$	50/50	28.0	20.0	3.470
Q35	38.0	-0.140	$[-0.350, +0.070]$	50/50	26.0	36.0	3.490
GLM	50.0	-0.240	$[-0.457, -0.023]$	50/50	24.0	26.0	3.150
DSK	30.0	-0.060	$[-0.311, +0.191]$	50/50	38.0	32.0	3.500
NMU	24.0	+0.010	$[-0.238, +0.258]$	47/50	46.0	30.0	2.936
GMA	18.0	+0.110	$[-0.010, +0.230]$	50/50	52.0	30.0	3.970
Q25	6.0	+0.125	$[-0.050, +0.300]$	49/50	80.0	14.0	3.265
MP14	12.0	-0.025	$[-0.326, +0.276]$	50/50	80.0	8.0	1.575
CEJ	12.0	+0.075	$[-0.225, +0.375]$	49/50	76.0	12.0	4.413
MP3	12.0	+0.100	$[-0.301, +0.501]$	50/50	68.0	20.0	1.250
NML	18.0	-0.070	$[-0.268, +0.128]$	50/50	74.0	8.0	3.040
MP7	6.0	+0.075	$[-0.116, +0.266]$	50/50	86.0	8.0	3.075
Subtle translationese							
FAB	0.0	+0.518	$[+0.326, +0.710]$	28/28	35.7	64.3	4.339
GEM	0.0	+0.720	$[+0.536, +0.904]$	50/50	26.0	74.0	4.510
JEV	4.0	+0.523	$[+0.373, +0.674]$	50/50	0.0	96.0	4.071
GPT	0.0	+0.620	$[+0.484, +0.756]$	50/50	18.0	82.0	4.320
Q35	4.0	+0.680	$[+0.513, +0.847]$	50/50	22.0	74.0	4.310
GLM	4.0	+0.860	$[+0.670, +1.050]$	50/50	18.0	78.0	4.250
DSK	6.0	+0.830	$[+0.577, +1.083]$	50/50	30.0	64.0	4.390
NMU	<u>12.0</u>	+0.680	$[+0.384, +0.976]$	47/50	30.0	58.0	3.649
GMA	0.0	+0.340	$[+0.226, +0.454]$	50/50	50.0	50.0	4.200
Q25	10.0	+0.050	$[-0.221, +0.321]$	49/50	74.0	16.0	3.189
MP14	2.0	+0.300	$[+0.052, +0.548]$	50/50	82.0	16.0	1.900
CEJ	20.0	+0.200	$[-0.116, +0.516]$	49/50	54.0	26.0	4.541
MP3	6.0	+0.300	$[-0.032, +0.632]$	50/50	66.0	28.0	1.450
NML	6.0	+0.350	$[+0.127, +0.573]$	50/50	64.0	30.0	3.460
MP7	2.0	+0.775	$[+0.392, +1.158]$	50/50	68.0	30.0	3.775

A.3.19 Scottish Gaelic

Table 23: Scottish Gaelic: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	14.0	+0.116	$[-0.065, +0.297]$	43/43	37.2	48.8	2.430
GEM	28.0	+0.030	$[-0.180, +0.240]$	50/50	36.0	36.0	2.650
JEV	30.0	+0.114	$[+0.014, +0.214]$	50/50	2.0	68.0	3.192
GPT	16.0	+0.020	$[-0.138, +0.178]$	50/50	62.0	22.0	3.040
Q35	26.0	+0.010	$[-0.204, +0.224]$	50/50	32.0	42.0	2.790
GLM	22.0	+0.070	$[-0.128, +0.268]$	50/50	40.0	38.0	2.870
DSK	20.0	+0.120	$[-0.142, +0.382]$	50/50	44.0	36.0	2.930
NMU	20.0	+0.030	$[-0.096, +0.156]$	50/50	62.0	18.0	0.210
GMA	10.0	+0.120	$[+0.021, +0.219]$	50/50	56.0	34.0	3.390
Q25	28.0	-0.300	$[-0.647, +0.047]$	50/50	54.0	18.0	3.125
MP14	14.0	+0.050	$[-0.293, +0.393]$	50/50	68.0	18.0	2.775
CEJ	24.0	-0.200	$[-0.586, +0.186]$	50/50	60.0	16.0	4.200
MP3	12.0	-0.100	$[-0.599, +0.399]$	50/50	78.0	10.0	2.825
NML	14.0	+0.090	$[-0.218, +0.398]$	50/50	72.0	14.0	2.680
MP7	8.0	-0.050	$[-0.170, +0.070]$	49/50	88.0	4.0	4.439
Mild translationese							
FAB	18.6	+0.081	$[-0.160, +0.323]$	43/43	34.9	46.5	2.395
GEM	26.0	+0.130	$[-0.077, +0.337]$	50/50	28.0	46.0	2.750
JEV	30.0	+0.173	$[+0.053, +0.293]$	50/50	0.0	70.0	3.251
GPT	14.0	+0.150	$[-0.011, +0.311]$	50/50	46.0	40.0	3.170
Q35	20.0	+0.140	$[-0.096, +0.376]$	50/50	40.0	40.0	2.920
GLM	16.0	+0.170	$[-0.053, +0.393]$	50/50	40.0	44.0	2.970
DSK	16.0	+0.320	$[+0.010, +0.630]$	50/50	40.0	44.0	3.130
NMU	22.0	-0.030	$[-0.143, +0.083]$	50/50	58.0	20.0	0.150
GMA	8.0	+0.170	$[+0.059, +0.281]$	50/50	54.0	38.0	3.440
Q25	36.0	-0.200	$[-0.604, +0.204]$	50/50	38.0	26.0	3.225
MP14	14.0	+0.100	$[-0.276, +0.476]$	50/50	64.0	22.0	2.825
CEJ	14.0	+0.025	$[-0.392, +0.442]$	50/50	72.0	14.0	4.425
MP3	12.0	-0.175	$[-0.589, +0.239]$	50/50	82.0	6.0	2.750
NML	12.0	-0.010	$[-0.343, +0.323]$	50/50	74.0	14.0	2.580
MP7	6.0	+0.075	$[-0.088, +0.238]$	49/50	84.0	10.0	4.566
Subtle translationese							
FAB	11.6	+0.570	$[+0.290, +0.850]$	43/43	16.3	72.1	2.884
GEM	2.0	+0.600	$[+0.408, +0.792]$	50/50	32.0	66.0	3.220
JEV	12.0	+0.408	$[+0.278, +0.539]$	50/50	2.0	86.0	3.486
GPT	2.0	+0.410	$[+0.260, +0.560]$	50/50	36.0	62.0	3.430
Q35	8.0	+0.440	$[+0.195, +0.685]$	50/50	30.0	62.0	3.220
GLM	4.0	+0.580	$[+0.366, +0.794]$	50/50	30.0	66.0	3.380
DSK	10.0	+0.690	$[+0.414, +0.966]$	50/50	34.0	56.0	3.500
NMU	22.0	-0.040	$[-0.151, +0.071]$	50/50	66.0	12.0	0.140
GMA	4.0	+0.320	$[+0.212, +0.428]$	50/50	36.0	60.0	3.590
Q25	38.0	-0.300	$[-0.687, +0.087]$	50/50	34.0	28.0	3.125
MP14	16.0	+0.050	$[-0.307, +0.407]$	50/50	64.0	20.0	2.775
CEJ	18.0	-0.225	$[-0.606, +0.156]$	50/50	66.0	16.0	4.175
MP3	8.0	+0.175	$[-0.245, +0.595]$	50/50	76.0	16.0	3.100
NML	16.0	+0.090	$[-0.181, +0.361]$	50/50	70.0	14.0	2.680
MP7	10.0	-0.075	$[-0.278, +0.128]$	49/50	82.0	8.0	4.413

A.3.20 Slovenian

Table 24: Slovenian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	92.9	<u>-1.798</u>	[-2.085, -1.510]	42/42	4.8	2.4	1.869
GEM	92.0	-1.830	[-2.097, -1.563]	50/50	6.0	2.0	2.020
JEV	<u>92.0</u>	-0.798	[-0.943, -0.653]	50/50	2.0	6.0	2.988
GPT	90.0	-1.300	[-1.488, -1.112]	50/50	8.0	2.0	2.640
Q35	80.0	-1.160	[-1.400, -0.920]	50/50	14.0	6.0	2.710
GLM	86.0	-1.130	[-1.343, -0.917]	50/50	8.0	6.0	2.490
DSK	74.0	-1.310	[-1.635, -0.985]	50/50	16.0	10.0	2.530
NMU	22.0	+0.230	[-0.067, +0.527]	48/50	30.0	48.0	1.094
GMA	66.0	-0.390	[-0.499, -0.281]	50/50	32.0	2.0	3.730
Q25	12.0	+0.025	[-0.153, +0.203]	50/50	74.0	14.0	4.000
MP14	22.0	-0.450	[-0.857, -0.043]	50/50	70.0	8.0	1.875
CEJ	12.0	+0.100	[-0.141, +0.341]	50/50	74.0	14.0	4.675
MP3	18.0	+0.050	[-0.376, +0.476]	50/50	64.0	18.0	2.250
NML	6.0	+0.110	[-0.061, +0.281]	50/50	82.0	12.0	2.920
MP7	6.0	+0.075	[-0.172, +0.322]	50/50	84.0	10.0	4.450
Mild translationese							
FAB	64.3	-0.548	[-0.806, -0.289]	42/42	14.3	21.4	3.119
GEM	62.0	-0.450	[-0.662, -0.238]	50/50	18.0	20.0	3.400
JEV	54.0	+0.096	[-0.044, +0.235]	50/50	0.0	46.0	3.883
GPT	<u>64.0</u>	-0.410	[-0.565, -0.255]	50/50	22.0	14.0	3.530
Q35	42.0	-0.180	[-0.395, +0.035]	50/50	34.0	24.0	3.690
GLM	42.0	-0.220	[-0.397, -0.043]	50/50	38.0	20.0	3.400
DSK	34.0	-0.150	[-0.402, +0.102]	50/50	36.0	30.0	3.690
NMU	28.0	+0.190	[-0.106, +0.486]	48/50	30.0	42.0	1.052
GMA	10.0	+0.130	[+0.019, +0.241]	50/50	64.0	26.0	4.250
Q25	8.0	+0.100	[-0.069, +0.269]	50/50	76.0	16.0	4.075
MP14	0.0	+0.300	[+0.062, +0.538]	50/50	86.0	14.0	2.625
CEJ	10.0	+0.175	[-0.046, +0.396]	50/50	70.0	20.0	4.750
MP3	0.0	+0.775	[+0.374, +1.176]	50/50	72.0	28.0	2.975
NML	4.0	+0.150	[+0.001, +0.299]	50/50	80.0	16.0	2.960
MP7	12.0	+0.075	[-0.161, +0.311]	50/50	76.0	12.0	4.450
Subtle translationese							
FAB	4.8	+0.524	[+0.272, +0.776]	42/42	47.6	47.6	4.190
GEM	0.0	+0.750	[+0.533, +0.967]	50/50	32.0	68.0	4.600
JEV	<u>18.0</u>	+0.379	[+0.216, +0.543]	50/50	2.0	80.0	4.166
GPT	2.0	+0.470	[+0.311, +0.629]	50/50	38.0	60.0	4.410
Q35	12.0	+0.590	[+0.383, +0.797]	50/50	24.0	64.0	4.460
GLM	6.0	+0.690	[+0.505, +0.875]	50/50	24.0	70.0	4.310
DSK	2.0	+0.850	[+0.580, +1.120]	50/50	42.0	56.0	4.690
NMU	38.0	-0.120	[-0.347, +0.107]	48/50	34.0	28.0	0.729
GMA	2.0	+0.340	[+0.196, +0.484]	50/50	54.0	44.0	4.460
Q25	4.0	+0.100	[-0.037, +0.237]	50/50	84.0	12.0	4.075
MP14	2.0	+0.450	[+0.068, +0.832]	50/50	78.0	20.0	2.775
CEJ	8.0	+0.075	[-0.088, +0.238]	50/50	78.0	14.0	4.650
MP3	4.0	+0.700	[+0.251, +1.149]	50/50	74.0	22.0	2.900
NML	8.0	+0.130	[-0.077, +0.337]	50/50	76.0	16.0	2.940
MP7	8.0	+0.175	[-0.057, +0.407]	50/50	76.0	16.0	4.550

A.3.21 Spanish

Table 25: Spanish: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	<u>97.7</u>	-2.012	[-2.229, -1.794]	43/43	0.0	2.3	1.860
GEM	98.0	<u>-1.760</u>	[-1.952, -1.568]	50/50	2.0	0.0	2.320
JEV	98.0	<u>-1.025</u>	[-1.167, -0.883]	50/50	0.0	2.0	2.884
GPT	98.0	-1.240	[-1.396, -1.084]	50/50	0.0	2.0	2.780
Q35	78.0	-0.970	[-1.202, -0.738]	50/50	16.0	6.0	2.880
GLM	88.0	-1.220	[-1.434, -1.006]	50/50	10.0	2.0	2.550
DSK	92.0	-1.430	[-1.674, -1.186]	50/50	4.0	4.0	2.600
NMU	46.0	-0.420	[-0.611, -0.229]	47/50	42.0	12.0	2.532
GMA	84.0	-0.780	[-0.920, -0.640]	50/50	16.0	0.0	3.370
Q25	6.0	-0.050	[-0.148, +0.048]	50/50	92.0	2.0	3.550
MP14	40.0	-0.925	[-1.314, -0.536]	50/50	58.0	2.0	1.425
CEJ	20.0	-0.100	[-0.296, +0.096]	50/50	68.0	12.0	4.600
MP3	44.0	-0.825	[-1.165, -0.485]	50/50	56.0	0.0	0.825
NML	18.0	-0.230	[-0.381, -0.079]	50/50	82.0	0.0	2.970
MP7	34.0	-0.725	[-1.103, -0.347]	50/50	60.0	6.0	3.200
Mild translationese							
FAB	88.4	-0.965	[-1.175, -0.755]	43/43	7.0	4.7	2.907
GEM	78.0	-0.830	[-1.033, -0.627]	50/50	18.0	4.0	3.250
JEV	68.0	-0.125	[-0.224, -0.027]	50/50	2.0	30.0	3.784
GPT	<u>86.0</u>	-0.690	[-0.818, -0.562]	50/50	12.0	2.0	3.330
Q35	54.0	-0.390	[-0.585, -0.195]	50/50	28.0	18.0	3.460
GLM	66.0	-0.680	[-0.861, -0.499]	50/50	30.0	4.0	3.090
DSK	54.0	-0.460	[-0.724, -0.196]	50/50	26.0	20.0	3.570
NMU	22.0	+0.040	[-0.118, +0.198]	47/50	50.0	28.0	3.021
GMA	34.0	-0.130	[-0.234, -0.026]	50/50	56.0	10.0	4.020
Q25	0.0	+0.125	[+0.020, +0.230]	50/50	90.0	10.0	3.725
MP14	18.0	-0.025	[-0.357, +0.307]	50/50	64.0	18.0	2.325
CEJ	10.0	+0.000	[-0.157, +0.157]	50/50	80.0	10.0	4.700
MP3	18.0	-0.300	[-0.538, -0.062]	50/50	80.0	2.0	1.350
NML	14.0	-0.130	[-0.290, +0.030]	50/50	82.0	4.0	3.070
MP7	10.0	-0.100	[-0.379, +0.179]	50/50	82.0	8.0	3.825
Subtle translationese							
FAB	4.7	+0.442	[+0.275, +0.609]	43/43	30.2	65.1	4.314
GEM	<u>8.0</u>	+0.540	[+0.398, +0.682]	50/50	18.0	74.0	4.620
JEV	10.0	+0.283	[+0.195, +0.371]	50/50	4.0	86.0	4.192
GPT	2.0	+0.400	[+0.299, +0.501]	50/50	30.0	68.0	4.420
Q35	6.0	+0.550	[+0.381, +0.719]	50/50	30.0	64.0	4.400
GLM	<u>8.0</u>	+0.480	[+0.329, +0.631]	50/50	28.0	64.0	4.250
DSK	2.0	+0.660	[+0.452, +0.868]	50/50	40.0	58.0	4.690
NMU	10.0	+0.570	[+0.312, +0.828]	47/50	38.0	52.0	3.585
GMA	0.0	+0.250	[+0.165, +0.335]	50/50	56.0	44.0	4.400
Q25	4.0	+0.025	[-0.085, +0.135]	50/50	90.0	6.0	3.625
MP14	4.0	+0.575	[+0.260, +0.890]	50/50	64.0	32.0	2.925
CEJ	4.0	+0.225	[+0.019, +0.431]	50/50	78.0	18.0	4.925
MP3	2.0	+0.150	[-0.067, +0.367]	50/50	92.0	6.0	1.800
NML	<u>8.0</u>	+0.280	[+0.068, +0.492]	50/50	64.0	28.0	3.480
MP7	4.0	+0.300	[-0.025, +0.625]	50/50	84.0	12.0	4.225

A.3.22 Swedish

Table 26: Swedish: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	95.0	-1.863	[-2.141, -1.584]	40/40	2.5	2.5	1.788
GEM	90.0	-1.550	[-1.816, -1.284]	50/50	4.0	6.0	2.110
JEV	88.0	-0.732	[-0.930, -0.534]	50/50	0.0	12.0	3.007
GPT	90.0	-1.130	[-1.324, -0.936]	50/50	6.0	4.0	2.610
Q35	76.0	-1.080	[-1.345, -0.815]	50/50	14.0	10.0	2.690
GLM	80.0	-1.210	[-1.468, -0.952]	50/50	12.0	8.0	2.380
DSK	78.0	-1.240	[-1.541, -0.939]	50/50	16.0	6.0	2.430
NMU	28.0	+0.020	[-0.195, +0.235]	40/50	34.0	38.0	1.812
GMA	76.0	-0.530	[-0.692, -0.368]	50/50	18.0	6.0	3.440
Q25	2.0	+0.125	[-0.050, +0.300]	50/50	90.0	8.0	3.325
MP14	34.0	-0.425	[-0.674, -0.176]	50/50	60.0	6.0	1.550
CEJ	12.0	+0.175	[-0.138, +0.488]	50/50	70.0	18.0	4.475
MP3	12.0	-0.250	[-0.571, +0.071]	50/50	84.0	4.0	1.550
NML	18.0	-0.170	[-0.337, -0.003]	50/50	80.0	2.0	2.920
MP7	34.0	-0.275	[-0.613, +0.063]	50/50	50.0	16.0	3.625
Mild translationese							
FAB	80.0	-0.912	[-1.184, -0.641]	40/40	7.5	12.5	2.737
GEM	70.0	-0.720	[-0.956, -0.484]	50/50	14.0	16.0	2.940
JEV	56.0	+0.024	[-0.090, +0.137]	50/50	2.0	42.0	3.763
GPT	70.0	-0.530	[-0.704, -0.356]	50/50	22.0	8.0	3.210
Q35	50.0	-0.410	[-0.640, -0.180]	50/50	36.0	14.0	3.360
GLM	60.0	-0.580	[-0.798, -0.362]	50/50	26.0	14.0	3.010
DSK	42.0	-0.280	[-0.573, +0.013]	50/50	26.0	32.0	3.390
NMU	16.0	+0.350	[+0.102, +0.598]	40/50	42.0	42.0	2.225
GMA	22.0	+0.040	[-0.091, +0.171]	50/50	60.0	18.0	4.010
Q25	4.0	+0.250	[+0.052, +0.448]	50/50	76.0	20.0	3.450
MP14	12.0	+0.000	[-0.210, +0.210]	50/50	80.0	8.0	1.975
CEJ	16.0	+0.050	[-0.202, +0.302]	50/50	66.0	18.0	4.350
MP3	14.0	+0.150	[-0.244, +0.544]	50/50	74.0	12.0	1.950
NML	12.0	+0.010	[-0.170, +0.190]	50/50	78.0	10.0	3.100
MP7	20.0	+0.125	[-0.157, +0.407]	50/50	56.0	24.0	4.025
Subtle translationese							
FAB	15.0	+0.425	[+0.210, +0.640]	40/40	35.0	50.0	4.075
GEM	4.0	+0.520	[+0.375, +0.665]	50/50	28.0	68.0	4.180
JEV	20.0	+0.295	[+0.177, +0.413]	50/50	0.0	80.0	4.034
GPT	2.0	+0.350	[+0.241, +0.459]	50/50	40.0	58.0	4.090
Q35	10.0	+0.350	[+0.193, +0.507]	50/50	32.0	58.0	4.120
GLM	4.0	+0.360	[+0.232, +0.488]	50/50	46.0	50.0	3.950
DSK	10.0	+0.620	[+0.400, +0.840]	50/50	28.0	62.0	4.290
NMU	16.0	+0.360	[+0.129, +0.591]	40/50	40.0	44.0	2.237
GMA	4.0	+0.210	[+0.081, +0.339]	50/50	66.0	30.0	4.180
Q25	8.0	+0.200	[-0.046, +0.446]	50/50	76.0	16.0	3.400
MP14	0.0	+0.375	[+0.162, +0.588]	50/50	76.0	24.0	2.350
CEJ	10.0	+0.225	[-0.024, +0.474]	50/50	68.0	22.0	4.525
MP3	4.0	+0.275	[-0.048, +0.598]	50/50	80.0	16.0	2.075
NML	6.0	+0.480	[+0.236, +0.724]	50/50	64.0	30.0	3.570
MP7	10.0	+0.225	[-0.024, +0.474]	50/50	68.0	22.0	4.125

A.3.23 Ukrainian

Table 27: Ukrainian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	97.8	-2.011	$[-2.273, -1.750]$	45/45	2.2	0.0	1.733
GEM	94.0	-1.940	$[-2.204, -1.676]$	50/50	6.0	0.0	2.010
JEV	<u>96.0</u>	-1.133	$[-1.297, -0.970]$	50/50	0.0	4.0	2.952
GPT	<u>96.0</u>	-1.410	$[-1.599, -1.221]$	50/50	4.0	0.0	2.550
Q35	92.0	-1.470	$[-1.733, -1.207]$	50/50	6.0	2.0	2.590
GLM	<u>96.0</u>	-1.430	$[-1.638, -1.222]$	50/50	2.0	2.0	2.350
DSK	88.0	-1.640	$[-1.917, -1.363]$	50/50	8.0	4.0	2.500
NMU	46.0	-0.170	$[-0.459, +0.119]$	48/50	26.0	28.0	1.562
GMA	84.0	-0.670	$[-0.803, -0.537]$	50/50	14.0	2.0	3.550
Q25	16.0	-0.025	$[-0.217, +0.167]$	50/50	70.0	14.0	4.025
MP14	20.0	-0.475	$[-0.838, -0.112]$	49/50	74.0	6.0	1.148
CEJ	14.0	+0.125	$[-0.174, +0.424]$	50/50	70.0	16.0	4.450
MP3	36.0	-0.575	$[-0.984, -0.166]$	50/50	56.0	8.0	1.425
NML	2.0	+0.060	$[-0.061, +0.181]$	50/50	94.0	4.0	3.000
MP7	16.0	-0.250	$[-0.614, +0.114]$	50/50	68.0	16.0	3.775
Mild translationese							
FAB	73.3	-0.700	$[-0.974, -0.426]$	45/45	11.1	15.6	3.044
GEM	68.0	-0.660	$[-0.939, -0.381]$	50/50	14.0	18.0	3.290
JEV	58.0	-0.002	$[-0.119, +0.114]$	50/50	0.0	42.0	4.083
GPT	<u>70.0</u>	-0.560	$[-0.749, -0.371]$	50/50	16.0	14.0	3.400
Q35	60.0	-0.540	$[-0.807, -0.273]$	50/50	22.0	18.0	3.520
GLM	<u>70.0</u>	-0.700	$[-0.910, -0.490]$	50/50	16.0	14.0	3.080
DSK	54.0	-0.420	$[-0.767, -0.073]$	50/50	24.0	22.0	3.720
NMU	28.0	+0.140	$[-0.147, +0.427]$	48/50	28.0	44.0	1.885
GMA	14.0	-0.020	$[-0.088, +0.048]$	50/50	76.0	10.0	4.200
Q25	20.0	-0.050	$[-0.282, +0.182]$	50/50	62.0	18.0	4.000
MP14	12.0	-0.225	$[-0.474, +0.024]$	49/50	84.0	4.0	1.403
CEJ	4.0	+0.300	$[+0.042, +0.558]$	50/50	78.0	18.0	4.625
MP3	10.0	+0.125	$[-0.234, +0.484]$	50/50	76.0	14.0	2.125
NML	4.0	+0.190	$[+0.005, +0.375]$	50/50	86.0	10.0	3.130
MP7	6.0	+0.250	$[-0.012, +0.512]$	50/50	78.0	16.0	4.275
Subtle translationese							
FAB	4.4	+0.722	$[+0.501, +0.944]$	45/45	31.1	64.4	4.467
GEM	4.0	+0.800	$[+0.573, +1.027]$	50/50	28.0	68.0	4.750
JEV	<u>12.0</u>	+0.315	$[+0.213, +0.418]$	50/50	0.0	88.0	4.401
GPT	0.0	+0.550	$[+0.401, +0.699]$	50/50	32.0	68.0	4.510
Q35	6.0	+0.620	$[+0.429, +0.811]$	50/50	30.0	64.0	4.680
GLM	4.0	+0.690	$[+0.505, +0.875]$	50/50	26.0	70.0	4.470
DSK	6.0	+0.590	$[+0.318, +0.862]$	50/50	44.0	50.0	4.730
NMU	18.0	+0.350	$[+0.078, +0.622]$	48/50	28.0	54.0	2.104
GMA	0.0	+0.210	$[+0.126, +0.294]$	50/50	62.0	38.0	4.430
Q25	<u>12.0</u>	-0.050	$[-0.271, +0.171]$	50/50	76.0	12.0	4.000
MP14	4.0	+0.525	$[+0.167, +0.883]$	49/50	74.0	22.0	2.168
CEJ	4.0	+0.325	$[+0.056, +0.594]$	50/50	78.0	18.0	4.650
MP3	6.0	+0.525	$[+0.141, +0.909]$	50/50	70.0	24.0	2.525
NML	8.0	+0.080	$[-0.058, +0.218]$	50/50	82.0	10.0	3.020
MP7	8.0	+0.275	$[-0.008, +0.558]$	50/50	74.0	18.0	4.300

A.3.24 Welsh

Table 28: Welsh: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	71.4	-1.071	$[-1.488, -0.655]$	42/42	19.0	9.5	2.917
GEM	56.0	-0.940	$[-1.389, -0.491]$	50/50	32.0	12.0	3.430
JEV	50.0	-0.222	$[-0.415, -0.030]$	50/50	8.0	42.0	3.874
GPT	50.0	-0.410	$[-0.721, -0.099]$	50/50	32.0	18.0	3.670
Q35	54.0	-0.870	$[-1.335, -0.405]$	50/50	24.0	22.0	3.170
GLM	50.0	-0.410	$[-0.718, -0.102]$	50/50	26.0	24.0	3.810
DSK	54.0	-1.060	$[-1.565, -0.555]$	50/50	30.0	16.0	3.250
NMU	12.0	+0.000	$[-0.093, +0.093]$	49/50	74.0	14.0	0.092
GMA	36.0	-0.520	$[-0.882, -0.158]$	50/50	44.0	20.0	3.430
Q25	28.0	-0.300	$[-0.585, -0.015]$	50/50	62.0	10.0	2.875
MP14	30.0	-0.550	$[-0.988, -0.112]$	50/50	56.0	14.0	2.950
CEJ	28.0	-0.350	$[-0.830, +0.130]$	50/50	54.0	18.0	4.225
MP3	12.0	+0.100	$[-0.428, +0.628]$	50/50	74.0	14.0	3.450
NML	20.0	-0.540	$[-0.880, -0.200]$	50/50	80.0	0.0	2.440
MP7	18.0	-0.400	$[-0.799, -0.001]$	50/50	68.0	14.0	3.975
Mild translationese							
FAB	59.5	-0.964	$[-1.436, -0.493]$	42/42	21.4	19.0	3.024
GEM	52.0	-0.930	$[-1.421, -0.439]$	50/50	26.0	22.0	3.440
JEV	44.0	-0.130	$[-0.307, +0.048]$	50/50	2.0	54.0	3.967
GPT	34.0	-0.190	$[-0.513, +0.133]$	50/50	40.0	26.0	3.890
Q35	44.0	-0.620	$[-1.110, -0.130]$	50/50	24.0	32.0	3.420
GLM	28.0	-0.140	$[-0.484, +0.204]$	50/50	44.0	28.0	4.080
DSK	40.0	-0.650	$[-1.174, -0.126]$	50/50	36.0	24.0	3.660
NMU	12.0	+0.010	$[-0.076, +0.096]$	49/50	74.0	14.0	0.102
GMA	28.0	-0.530	$[-0.919, -0.141]$	50/50	48.0	24.0	3.420
Q25	24.0	-0.300	$[-0.632, +0.032]$	50/50	66.0	10.0	2.875
MP14	32.0	-0.675	$[-1.090, -0.260]$	50/50	58.0	10.0	2.825
CEJ	16.0	-0.350	$[-0.830, +0.130]$	50/50	64.0	20.0	4.225
MP3	20.0	-0.025	$[-0.587, +0.537]$	50/50	64.0	16.0	3.325
NML	18.0	-0.580	$[-0.935, -0.225]$	50/50	82.0	0.0	2.400
MP7	24.0	-0.575	$[-0.972, -0.178]$	50/50	68.0	8.0	3.800
Subtle translationese							
FAB	35.7	-0.655	$[-1.165, -0.144]$	42/42	40.5	23.8	3.333
GEM	22.0	-0.470	$[-0.959, +0.019]$	50/50	38.0	40.0	3.900
JEV	44.0	-0.114	$[-0.296, +0.068]$	50/50	0.0	56.0	3.982
GPT	12.0	+0.120	$[-0.218, +0.458]$	50/50	38.0	50.0	4.200
Q35	38.0	-0.540	$[-1.086, +0.006]$	50/50	22.0	40.0	3.500
GLM	16.0	-0.050	$[-0.433, +0.333]$	50/50	40.0	44.0	4.170
DSK	24.0	-0.300	$[-0.820, +0.220]$	50/50	50.0	26.0	4.010
NMU	10.0	+0.030	$[-0.056, +0.116]$	49/50	74.0	16.0	0.122
GMA	24.0	-0.400	$[-0.766, -0.034]$	50/50	44.0	32.0	3.550
Q25	24.0	-0.200	$[-0.531, +0.131]$	50/50	60.0	16.0	2.975
MP14	30.0	-0.500	$[-0.937, -0.063]$	50/50	50.0	20.0	3.000
CEJ	24.0	-0.375	$[-0.802, +0.052]$	50/50	60.0	16.0	4.200
MP3	14.0	+0.025	$[-0.472, +0.522]$	50/50	70.0	16.0	3.375
NML	20.0	-0.560	$[-0.928, -0.192]$	50/50	78.0	2.0	2.420
MP7	24.0	-0.550	$[-0.947, -0.153]$	50/50	66.0	10.0	3.825

A.3.25 Western Frisian

Table 29: Western Frisian: complete results in heavy, mild, and subtle blocks. Models follow overall average detection order within each block. Rates are percentages.

Model	Detect $\uparrow$	Δ $\downarrow$	95% CI, Δ	Scored/ N	Ties	Reverse	Mean score
Heavy translationese							
FAB	65.9	-1.049	$[-1.518, -0.580]$	41/41	19.5	14.6	2.780
GEM	64.0	-0.980	$[-1.385, -0.575]$	50/50	20.0	16.0	3.230
JEV	42.0	+0.022	$[-0.108, +0.152]$	50/50	2.0	56.0	3.732
GPT	50.0	-0.340	$[-0.539, -0.141]$	50/50	38.0	12.0	3.650
Q35	50.0	-0.540	$[-1.016, -0.064]$	50/50	10.0	40.0	3.160
GLM	44.0	-0.070	$[-0.331, +0.191]$	50/50	18.0	38.0	3.630
DSK	36.0	-0.370	$[-0.885, +0.145]$	50/50	20.0	44.0	3.320
NMU	48.0	-0.294	$[-0.697, +0.109]$	50/50	6.0	46.0	1.476
GMA	24.0	-0.320	$[-0.556, -0.084]$	50/50	64.0	12.0	3.970
Q25	26.0	-0.050	$[-0.503, +0.403]$	46/50	48.0	26.0	3.451
MP14	30.0	-0.300	$[-0.632, +0.032]$	50/50	62.0	8.0	1.575
CEJ	10.0	+0.325	$[+0.013, +0.637]$	50/50	62.0	28.0	4.400
MP3	12.0	+0.625	$[+0.082, +1.168]$	49/50	56.0	32.0	1.658
NML	24.0	-0.010	$[-0.463, +0.443]$	50/50	50.0	26.0	2.240
MP7	12.0	-0.075	$[-0.252, +0.102]$	50/50	80.0	8.0	4.600
Mild translationese							
FAB	56.1	-0.841	$[-1.315, -0.368]$	41/41	22.0	22.0	2.988
GEM	44.0	-0.650	$[-1.125, -0.175]$	50/50	24.0	32.0	3.560
JEV	28.0	+0.117	$[+0.002, +0.231]$	50/50	0.0	72.0	3.827
GPT	28.0	-0.040	$[-0.162, +0.082]$	50/50	48.0	24.0	3.950
Q35	34.0	-0.390	$[-0.882, +0.102]$	50/50	14.0	52.0	3.310
GLM	20.0	+0.190	$[-0.118, +0.498]$	50/50	28.0	52.0	3.890
DSK	32.0	-0.280	$[-0.834, +0.274]$	50/50	14.0	54.0	3.410
NMU	42.0	-0.130	$[-0.554, +0.294]$	50/50	14.0	44.0	1.640
GMA	22.0	-0.300	$[-0.561, -0.039]$	50/50	58.0	20.0	3.990
Q25	40.0	-0.375	$[-0.802, +0.052]$	46/50	40.0	20.0	3.098
MP14	20.0	-0.175	$[-0.552, +0.202]$	50/50	70.0	10.0	1.700
CEJ	16.0	+0.275	$[-0.110, +0.660]$	50/50	50.0	34.0	4.350
MP3	12.0	+0.725	$[+0.215, +1.235]$	49/50	52.0	36.0	1.760
NML	24.0	-0.020	$[-0.452, +0.412]$	50/50	48.0	28.0	2.230
MP7	10.0	-0.125	$[-0.338, +0.088]$	50/50	84.0	6.0	4.550
Subtle translationese							
FAB	24.4	-0.341	$[-0.868, +0.185]$	41/41	31.7	43.9	3.488
GEM	18.0	-0.070	$[-0.547, +0.407]$	50/50	28.0	54.0	4.140
JEV	12.0	+0.188	$[+0.061, +0.316]$	50/50	0.0	88.0	3.899
GPT	2.0	+0.420	$[+0.307, +0.533]$	50/50	34.0	64.0	4.410
Q35	22.0	+0.050	$[-0.446, +0.546]$	50/50	16.0	62.0	3.750
GLM	8.0	+0.580	$[+0.308, +0.852]$	50/50	20.0	72.0	4.280
DSK	22.0	+0.190	$[-0.355, +0.735]$	50/50	18.0	60.0	3.880
NMU	34.0	-0.050	$[-0.416, +0.316]$	50/50	22.0	44.0	1.720
GMA	20.0	-0.260	$[-0.559, +0.039]$	50/50	50.0	30.0	4.030
Q25	32.0	-0.175	$[-0.607, +0.257]$	46/50	40.0	28.0	3.315
MP14	18.0	-0.075	$[-0.421, +0.271]$	50/50	70.0	12.0	1.800
CEJ	16.0	+0.325	$[-0.037, +0.687]$	50/50	56.0	28.0	4.400
MP3	16.0	+0.500	$[-0.051, +1.051]$	49/50	48.0	36.0	1.531
NML	24.0	-0.044	$[-0.479, +0.391]$	50/50	48.0	28.0	2.206
MP7	16.0	-0.225	$[-0.443, -0.007]$	50/50	80.0	4.0	4.450