

Calibrating an Open Reading-Difficulty Scale without Human-Labeled Corpora: LLM Pairwise Judgments, Deterministic Features, and Their Composition

Jasper Vande Moortel · Orion-Solution · Version 0.1 (draft) · July 2026

Abstract

The Free Reading Standard (FRS) specifies an open, continuous scale for the technical reading difficulty of Dutch and English texts, calibrated from pairwise difficulty judgments produced by a large language model rather than from human-labeled corpora, which for Dutch do not exist in open form. This paper tests whether that calibration method works. Against the CLEAR corpus, whose 4,724 excerpts carry continuous difficulty scores derived from 111,347 human pairwise judgments, an LLM judge (gpt-5.4-mini, temperature 0, every pair judged in both presentation orders) reproduced the human ordering at Spearman $\rho = .69$ on a 120-text stratified sample; rewording the prompt to match the human raters' own question raised agreement to $\rho = .75$, identifying construct mismatch as a real but minority factor. Thirteen deterministic linguistic features with a small regressor predicted the human scores at cross-validated $\rho = .72$ on the full corpus, outperforming Flesch-Kincaid (.55), a Lexile-style two-variable baseline (.66), and the technical-prompt LLM judge (.69); the construct-matched judge's .75 exceeds the feature models in rank agreement but requires hundreds of model calls per text set and proves a noisier training signal. The critical composition test trained the feature regressor on LLM-derived targets from only 120 anchor texts and evaluated it against human scores on 4,604 held-out texts: $\rho = .63$, retaining 89% of the same-architecture full-data ceiling (ridge, $\rho = .71$; 88% of the overall GBM ceiling of .72), with the shortfall decomposed against that same-architecture ceiling as roughly three-quarters label noise and one-quarter sample size, and no degradation cliff down to 60 anchors. The results support a license-clean production recipe (deterministic features plus a ridge model fitted on LLM pairwise placements of open anchor texts) and quantify its current cost in precision. Dutch-language validation beyond pilot scale, human judge comparison, and off-distribution testing remain open.

Keywords: readability, text difficulty, Bradley-Terry, large language models, calibration, CLEAR corpus, Dutch

1. Introduction

The companion specification paper defines the Free Reading Standard: a four-dimension difficulty profile plus a continuous composite score on an anchored 0–100 scale, for Dutch and English texts, with every parameter published. Its calibration method rests on a wager: that pairwise difficulty judgments produced by a large language model, aggregated with a Bradley-Terry model and pinned to anchor texts, can substitute for the human judgment campaigns and proprietary reader-response datasets that built the incumbent scales. The wager is forced rather than chosen. Dutch possesses no open corpus of difficulty-labeled texts; the nearest human-judged resource in any language, the CLEAR corpus (Crossley et al., 2023, building on the crowdsourced readability work of De Clercq et al., 2014), is licensed for non-commercial use and covers English only. If an open standard is to exist at all, its calibration signal must come from somewhere reproducible, affordable, and license-clean.

This paper subjects that wager to the strongest test currently available. CLEAR's continuous difficulty scores, derived from 111,347 pairwise judgments by 1,116 raters, serve as the human gold standard. We ask four questions. **RQ1:** how well does an LLM judge, operating under the FRS pairwise protocol, reproduce the human difficulty ordering? **RQ2:** how much of the shortfall is attributable to construct mismatch between the FRS prompt (technical reading difficulty) and the question CLEAR's raters answered (ease for their students)? **RQ3:** how much per-text precision do the standard's deterministic linguistic features carry, relative to classic baselines and to the LLM judge? **RQ4:** does the composition at the heart of the FRS design (a small regressor over deterministic features, trained on LLM-derived targets) retain enough of that precision when no human labels enter training at all?

Two small Dutch pilot experiments, reported in the specification paper and summarized here as baseline rows, motivated the design: absolute rubric scoring by an LLM proved reliable but compressed twenty texts into a quarter of its nominal scale, while pairwise judgment of the same texts nearly halved ordering errors. The present experiments test whether those observations survive contact with a realistic, heterogeneous, human-scored corpus two orders of magnitude larger.

2. Materials

CLEAR corpus. 4,724 English excerpts (140–200 words, grades 3–12, ~even informative/literary split), each with a continuous Bradley-Terry easiness score derived from pairwise comparisons by raters the corpus authors describe as practicing teachers (Crossley et al., 2023); CC BY-NC-SA 4.0, used here for research validation only. A stratified sample of 120 excerpts (20 per difficulty sextile) served the judgment experiments; the full corpus served the feature experiments.

Dutch pilot set. The twenty Dutch seed texts with integer difficulty labels from the legacy FRS system, as described in the specification paper, including its documented label-provenance caveats.

Judge model. gpt-5.4-mini (version 2026-03-17) on Azure AI Foundry, temperature 0, maximum 20 completion tokens with retry-on-empty. The Dutch pilots used gpt-4o-mini; judge-model stability across versions is explicitly not established by this paper.

Feature extractor. frs_core v0, the standard's open layer-1 implementation: thirteen deterministic features spanning decodability (word length in letters and syllables, consonant-cluster and long-word proportions), lexical difficulty (SUBTLEX-based content-word log frequency, low-frequency proportion, moving-average type-token ratio; English frequencies from SUBTLEX-US, Brysbaert & New, 2009), syntactic complexity (sentence length, mean and maximal dependency length, content words per finite clause), and referential cohesion (adjacent-sentence lemma overlap, connective density). Same text in, same numbers out, always.

3. Method

Pairwise protocol (RQ1, RQ2). From the 120-text sample, each text was paired with four nearest neighbors in the human ranking and two long-range partners from other strata, yielding 481 unique pairs. Every pair was judged in both presentation orders (962 calls per condition; in the matched condition, two texts persistently errored at the endpoint, excluding four judgments, 0.4%), following the both-orders rule established in the Dutch pilot, where 16% of verdicts flipped with presentation order. Two prompt conditions were run: the FRS technical prompt ("which text is technically harder to read: decodability and reading fluency demand, not content") and a construct-matched variant mirroring CLEAR's instruction to its raters ("as an experienced teacher, which passage would be easier for your students to read and understand"), with polarity aligned at analysis time. Judgments were aggregated per condition with a Bradley-Terry maximum-likelihood fit; the resulting latent difficulty values were compared with CLEAR's human scores by rank correlation.

Feature models (RQ3). Ridge regression and gradient boosting over the thirteen features, predicting human easiness scores on all 4,724 excerpts under 5-fold cross-validation with identical folds across models. Two baselines on the same folds: the corpus's own Flesch-Kincaid grade level, and a two-variable model of mean log word frequency plus mean sentence length, the variable pair of the original Lexile specification equation (Stenner, 1996).

Composition (RQ4). The production scenario: ridge and gradient-boosting models trained on the 120 sampled texts using only the LLM-derived Bradley-Terry values as targets, then evaluated against human scores on the 4,604 texts the LLM never saw. Controls isolate the two costs at play: models trained on the same 120 texts with human targets (isolating label quality

at fixed sample size), and the full-corpus cross-validated ceiling (isolating sample size). A sensitivity analysis repeated the procedure with five random 60-text training subsamples per target type, and the full composition analysis was repeated with the construct-matched targets.

All raw judgments, per-text results, fitted values, and analysis code are published with the reference implementation; experiment artifact paths are listed in Appendix B.

4. Results

4.1 Baselines: the Dutch pilots

On twenty curated Dutch texts, absolute rubric scoring was highly repeatable ($ICC(1,1) = .97$ at production settings, $.99$ at temperature 0) and correlated with the reference labels at $\rho = .93$, but used only 2.00–4.57 of its nominal 0–12 range, and 24 of 190 pair orderings contradicted the labels. Pairwise judgment on the identical texts and model reached $\rho = .944$ with 13 violations. These numbers, from small and favorable material, set expectations that the following experiments correct.

4.2 RQ1: LLM pairwise judgment against the human gold standard

On the 120-text CLEAR sample, the technical-prompt Bradley-Terry ordering correlated with human scores at $\rho = .686$ (Pearson $r = .661$). Presentation-order self-consistency was 81.1%. Within single difficulty sextiles, correlations were weak and inconsistent (ρ between $-.20$ and $.40$, none significant): the judge resolves broad differences, not fine ones. The pilot's $\rho = .944$ does not generalize; on realistic heterogeneous material, LLM pairwise judgment is a coarse ordering signal.

4.3 RQ2: the construct discount

The construct-matched prompt raised agreement with the human scores to $\rho = .749$, recovering roughly a quarter of the distance between the technical-prompt result and the Dutch pilot figure, at the cost of lower self-consistency (76.8%) and markedly lower linear correlation with the human values ($r = .49$ versus $.66$). The two prompts' orderings correlate at $\rho = .84$. Construct mismatch is therefore real and measurable but explains a minority of RQ1's shortfall; the remainder is attributable to the difficulty of the material itself. Two consequences follow. Validation of a technical-reading instrument against a general-readability gold standard carries an inherent construct discount that must not be misread as pure method error. And prompt framing moves agreement by economically meaningful amounts, though §4.5 shows the direction of improvement depends on what the labels are used for.

4.4 RQ3: deterministic features carry the precision

Model (5-fold CV, n = 4,724)	Pearson r	Spearman ρ
frs_core features + gradient boosting	.712 $\pm$.013	.720 $\pm$.011
frs_core features + ridge	.695 $\pm$.016	.712 $\pm$.015
Lexile-style baseline (log frequency + sentence length)	.648 $\pm$.019	.660 $\pm$.019
Flesch-Kincaid	.517 $\pm$.009	.547 $\pm$.012

Thirteen deterministic features outperform both classic baselines and the technical-prompt LLM judge of §4.2, at zero inference cost and with exact reproducibility. For completeness: the construct-matched judge of §4.3 reaches $\rho = .749$, slightly above the feature models in rank agreement, but at roughly eight model calls per text and, as §4.5 shows, with score values too noisy to serve as a superior training signal. Five features carry 86% of the model's importance: content-word log frequency (dominant, consistent with the Dutch feature-selection evidence of Kleijn, 2018), mean sentence length, low-frequency word proportion, long-word proportion, and mean dependency length.

4.5 RQ4: the composition test

Training target (n = 120)	Model	Held-out ρ (n = 4,604)
LLM Bradley-Terry values, technical prompt	ridge	.634
LLM Bradley-Terry values, matched prompt	ridge	.619
Human scores (same-n control)	ridge	.694
Human scores (full-CV ridge ceiling, from §4.4)	ridge	.712
Human scores (full-CV ceiling, from §4.4)	GBM	.720

Trained without a single human label, the ridge regressor retains **89% of its own full-data ceiling** (.634 of .712; 88% of the overall GBM ceiling). Decomposed against that same-architecture ceiling, the 0.078 shortfall splits cleanly: about three-quarters is label quality (0.059 — the same-n human control loses only 0.019 to the ridge ceiling) and one-quarter is sample size. Halving the anchor set to 60 texts costs a further $\sim .04$ with no cliff. Two second-order findings: gradient boosting overfits at this training size and loses to ridge throughout, and the construct-matched targets, despite agreeing better with humans in rank, train a *worse* ridge model, because their score values are noisier (the value-level correlation of §4.3) and ridge fits values, not ranks. The production recipe that follows: deterministic features, ridge, technical-

prompt LLM targets, anchor sets of roughly a hundred texts or more, with improvement effort aimed at judge label quality (comparison density, stronger judges) rather than anchor volume.

5. Discussion

The wager holds, with a stated price. An open scale can be calibrated end-to-end without human-labeled corpora at $\rho \approx .63$ against human judgment, where the ceiling with thousands of human labels is $\rho \approx .72$ on the same corpus, and the human-derived gold standard itself is built from judgments whose raters agree imperfectly with each other (Crossley et al., 2023, report split-half reliability near .85 for the corpus's scores). Whether .63 suffices depends on use: for placing texts in broad difficulty regions with published uncertainty bands, it does; for fine within-band discrimination, it does not yet, and the standard's uncertainty reporting exists precisely to keep that limitation visible.

Division of labor is now empirical. The specification's layer design assigned precision to deterministic features and scale anchoring to LLM pairwise judgment on theoretical grounds. The results confirm both directions of the assignment: the judge cannot discriminate within narrow bands (§4.2) but the features can (§4.4); the features cannot place themselves on an anchored scale without targets, and the judge supplies serviceable ones (§4.5). This mirrors the wider automatic-readability literature, where hybrid feature-plus-neural systems dominate on small data (Lee et al., 2021; Martinc et al., 2021), relative framing proves easier than absolute scoring for both fine-tuned rankers (Lee & Lee, 2023) and prompted models (whose absolute estimates plateau near $r \approx .75$ against human judgment; Trott & Rivière, 2024).

Label quality is the frontier. Three-quarters of the composition shortfall is label noise. The measured levers: comparison density per anchor (the sparse design used ~8 comparisons per text against CLEAR's ~46), judge model strength, and ensemble aggregation across judges. Prompt construct matching, the intuitive lever, improves rank agreement but degrades value quality, a distinction that only surfaced because the composition test measures what the labels are for rather than what they correlate with.

License hygiene worked as designed. CLEAR functioned exclusively as a research-mode gold standard; nothing trained on it ships. The production path trains on LLM placements of open anchor texts, and this separation is what allows FRS to publish everything.

Construct validity remains the deepest open issue. The judge encodes an unknown blend of difficulty notions from its training data; matching its instruction to a human rater population changes its behavior measurably (§4.3). FRS's mitigation is anchoring plus per-language human validation; for Dutch, whose gold standard does not yet exist, that validation is the standard's largest outstanding debt.

6. Limitations

(1) The gold standard is English-only, single-corpus, and its own construct (teacher-judged ease for students) differs from FRS's technical construct by a measured but not fully separable margin. (2) Dutch results remain pilot-scale ($n = 20$, labels of uncertain provenance). (3) One judge model family at one point in time; judge drift across versions is unquantified. (4) Feature-model results are in-corpus cross-validation on 140–200-word excerpts, grades 3–12; off-distribution generalization and full-spectrum coverage (first word lists, academic prose) are untested — a standing weakness of evaluation practice in this field generally (Vajjala, 2021). (5) frs_core vo's Dutch lexical frequencies use a provisional substitute list pending restored access to SUBTLEX-NL, and concreteness features are unimplemented. (6) No reader-side outcomes appear anywhere in this paper; the conjoint ambition is future work.

7. Conclusion and next steps

The FRS calibration wager survives its first serious test at a known, decomposed cost in precision. The registered next steps: score the open Dutch anchor corpus (staatsexamen texts and the Wikipedia–Wikikids pairs) with the locked recipe and validate against a purpose-built multi-rater human Dutch set; quantify judge stability across model versions and comparison densities; and stress the feature models off-distribution, where the legacy system's random forest originally failed. Each ships with published data, per the standard's openness requirements.

References

- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.
- Crossley, S. A., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55, 491–507. <https://doi.org/10.3758/s13428-022-01802-x>
- De Clercq, O., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., & Macken, L. (2014). Using the crowd for readability prediction. *Natural Language Engineering*, 20(3), 293–325.
- Kleijn, S. (2018). *Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing* [Doctoral dissertation, Utrecht University]. LOT. ISBN 978-94-6093-277-9.
- Lee, B. W., Jang, Y. S., & Lee, J. H.-J. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *Proceedings of EMNLP 2021*. arXiv:2109.12258.

- Lee, B. W., & Lee, J. H.-J. (2023). Prompt-based learning for text readability assessment. *Findings of EACL 2023*. arXiv:2302.13139.
- Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1), 141–179.
- Stenner, A. J. (1996, February). *Measuring reading comprehension with the Lexile Framework*. Paper presented at the California Comparability Symposium. (ERIC ED435977)
- Trott, S., & Rivière, C. (2024). *Measuring and modifying the readability of English texts with GPT-4*. arXiv:2410.14028.
- Vajjala, S. (2021). *Trends, limitations and open challenges in automatic readability assessment research*. arXiv:2105.00973.

Entries whose bibliographic details or content rest partly on secondary sources are used for context only; all load-bearing empirical claims in this paper derive from the published experiment artifacts of Appendix B and the primary-read sources (Crossley et al., 2023; Stenner, 1996; Kleijn, 2018).

Appendix A. Judgment prompts (verbatim)

Technical prompt (English), RQ1/RQ4: system: "You compare two English texts on TECHNICAL reading difficulty (decodability and reading fluency demand, not content). Answer only 'A' or 'B': the harder text." User message: the two texts as Text A / Text B (truncated to 150 words each), then "Which text is harder? Answer only 'A' or 'B':".

Construct-matched prompt (English), RQ2: system: "You are an experienced teacher comparing two English text excerpts. Answer only 'A' or 'B': the passage that would be EASIER for your students to read and understand." User message: the two texts as Text A / Text B (truncated to 150 words each), then "Which passage is easier? Answer only 'A' or 'B':". Polarity aligned at analysis time.

Dutch pilot prompts appear verbatim in the specification paper's Appendix B.

Appendix B. Experiment artifacts

Experiment	Location
Dutch pilots (absolute + pairwise)	experiments/scoring_baseline/ (runs.jsonl, pairwise_runs.jsonl, REPORT.md)
CLEAR benchmark, technical + matched prompts	experiments/clear_benchmark/ (runs.jsonl, runs_matched.jsonl, results*.json, REPORT.md)
Feature extraction + models + composition	experiments/clear_features/ (features.parquet, fit_results.json, composition*.json, REPORT.md)

Judge: gpt-5.4-mini (2026-03-17), temperature 0. Feature extractor: frs_core v0 (commit 5bf5de2). All artifacts ship with the reference implementation.