Applied Language Sciences

RESEARCH Open Access

Same CEFR, Different Prompt: Instruction-Paraphrase Sensitivity in CEFR-Conditioned Language Model Writing

Michele Carlo[1]*

*Correspondence:
m.carlo@me.com

[1] Graduate School of Foreign Language Education and Research
Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan

Abstract

This study tests whether equivalent instructions alter the linguistic features of writing generated at Common European Framework of Reference for Languages (CEFR) levels. Four instruction-tuned language models each generated responses at four CEFR levels across ten writing tasks, five equivalent instruction wordings, and five repetitions per condition (N = 4,000 texts). Twenty prespecified linguistic measures described vocabulary, grammar, discourse, and stance in each generated text. Requested level accounted for 27.8% of variation in these measures, whereas wording accounted for 0.1%; dispersion did not differ significantly across wordings. Classifier accuracy changed by at most 0.6 percentage points for an unseen wording, compared with an 11–16-point loss for an unseen task. Task effects remained large after adjustment for response length. Within these five paraphrases, task choice mattered more than instruction wording, supporting wording robustness under the tested conditions without establishing that generated texts validly represent human proficiency.

Keywords: CEFR; large language models; prompt sensitivity; synthetic learner writing; measurement validity; L2 writing

Check for updates

GloDAL
Global Digital Applied Linguistics Association

© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

1. Introduction

Asking a large language model (LLM) to write "at CEFR B2 level" offers a simple way to generate apparently proficiency-appropriate text. The Common European Framework of Reference for Languages (CEFR) describes six proficiency levels, A1 – C2; B2 denotes upper-intermediate proficiency. Level-conditioned generation supports graded materials, assessment prototypes, synthetic learner corpora, teacher education, and simulated proficiency-stratified samples. Research designs, including the author's own, commonly fix one instruction sentence and analyze the resulting corpus under its requested level labels.

This practice assumes that the level instruction provides a stable operationalization: linguistic properties should reflect the requested level rather than incidental phrasing. Otherwise, studies using different sentences to request B2 may sample different distributions while reporting under the same label.

Prompt-sensitivity research makes that assumption questionable. Benchmark performance changes with example ordering, formatting, and instruction paraphrases (Lu et al., 2022; Mizrahi et al., 2024; Sclar et al., 2024), while successful performance need not demonstrate interpretation of the intended instruction meaning (Webson & Pavlick, 2022). Whether comparable instability affects proficiency-conditioned writing remains an empirical question.

Here the concern is reproducibility, not which prompt performs best. In conditioned generation, the instruction is the manipulation. Alternative wordings should yield comparable feature locations, within-condition variability, and learnable level-related patterns. Stability would support the generation procedure, not establish correspondence with human CEFR proficiency.

Three gaps motivate the study. Empirically, prior CEFR and prompting research does not establish the stability of proficiency-conditioned linguistic profiles under paraphrase (Malik et al., 2024; Mizrahi et al., 2024; Wisniewski, 2017). Methodologically, benchmark accuracy does not capture shifts in location, dispersion, or classifier generalization. Practically, researchers need evidence for deciding when one CEFR instruction is adequate and when several should be sampled. A fully crossed, 4,000-text design addresses these gaps through multivariate analyses, held-out-wording classification, and the Prompt Sensitivity Ratio (PSR), which describes wording-related displacement relative to adjacent requested levels. The contribution is to identify conditions and limits under which CEFR-conditioned generation is a reproducible research operation, without equating reproducibility with human proficiency validity.

2. Literature Review

2.1 CEFR Proficiency and Level-Conditioned Generation

The Common European Framework of Reference describes language proficiency through communicative activities and competences rather than through fixed inventories of linguistic forms (Council of Europe, 2020). Its levels nevertheless correlate with measurable textual properties, and criterial-feature research has sought observable correlates of level distinctions in learner writing (Hawkins & Buttery, 2010). These

correlates are empirical operationalizations, not definitions of the levels. Wisniewski's (2017) critical examination of empirical learner language and CEFR levels is important here: a framework for describing communicative proficiency does not by itself supply a universal linguistic signature against which every text can be checked. Feature selection and interpretation therefore require evidence from the population and writing conditions concerned.

Journal research on syntactic complexity illustrates this dependence on context. Khushik and Huhta (2020) examined EFL writing across A1, A2, and B1 in different first-language populations, making learner background relevant to the interpretation of level-related complexity. Their later Finnish-background study found that sentence- and T-unit-level measures most clearly separated A1 – A2, whereas clausal and phrasal measures were more informative at B1 – B2 (Khushik & Huhta, 2022). The two designs differ in population coverage and proficiency range; they should not be treated as interchangeable demonstrations of a single complexity ladder. Cross-sectional associations can help discriminate rated levels, but they do not alone show how the same learner develops or whether an index transfers unchanged to another population. For generation research, this distinction matters because requested labels and task conditions can be standardized much more readily than human proficiency.

Yamashita (2025) adds a longitudinal perspective across five CEFR levels and emphasizes a robust measurement design for studying syntactic development. This complements rather than replaces cross-sectional evidence: change within learners and differences among proficiency groups answer different questions. Task variation and computational measurement also affect what a complexity trajectory represents. Read together, these studies support using multiple interpretable indicators while resisting the assumption that all should increase monotonically or carry the same meaning in every context. In the present study, the twenty indicators characterize generated writing; they are not presented as a complete CEFR scale or as substitutes for human judgments of communicative performance.

Automated CEFR classification provides a related but distinct use of linguistic evidence. Vajjala and Rama (2018) compared monolingual and cross-lingual classification, while Kerz et al. (2021) investigated sequential complexity representations rather than text-average features. Large error-annotated, CEFR-labeled corpora such as Write & Improve support this work (Nicholls et al., 2024). Predictive success establishes that representations contain information associated with the available labels under a particular evaluation design. It does not establish universal feature thresholds or justify transferring human proficiency interpretations to synthetic texts. The learner-language journal evidence and computational classification evidence thus address complementary questions: what linguistic differences accompany rated proficiency, and how effectively those differences support prediction.

Whereas discriminative approaches infer proficiency from observed text, level-conditioned generation shifts the task to generative control: the CEFR level becomes an input constraint, and the model produces text intended to instantiate it. The methodological question is whether alternative formulations elicit comparable linguistic realizations. Malik et al. (2024) compare prompting and fine-tuning for proficiency control, documenting usable but imperfect control through prompting alone. Such generation can support

materials development, assessment prototyping, and synthetic training data, but those uses depend on the interpretation of the generated level labels.

The author's previous work found level-ordered linguistic structure whose strength differed across models (Carlo & Takeuchi, 2026a). Relative to human writing, conditioned outputs diverged in stance and engagement (Carlo & Takeuchi, 2026c) and register adaptation, particularly at higher levels (Carlo & Takeuchi, 2026b). Those studies used a single instruction formulation, leaving wording sensitivity untested.

2.2 Prompt Sensitivity in Large Language Models

Mizrahi et al. (2024), in Transactions of the Association for Computational Linguistics, show that instruction paraphrases can alter benchmark scores and model rankings. Multi-prompt evaluation therefore addresses reproducibility, not merely prompt optimization. Their intervention differs from example ordering (Lu et al., 2022) and incidental formatting (Sclar et al., 2024). These manipulations should not be conflated: the present study isolates the proficiency sentence while fixing the surrounding task and instruction frame. While prompt-sensitivity research shows that benchmark scores and model outputs vary with prompt formatting and paraphrase (Mizrahi et al., 2024; Sclar et al., 2024), the strength and phrasing of revision instructions also heavily alter text profiles. For instance, Watson Todd (2026) demonstrated that varying prompt instructions from light surface editing to heavy revision drastically changes the linguistic profile of generated academic texts, shifting original wording retention from approximately 84% down to 18% and substantially increasing the frequency of rare lexical items.

Webson and Pavlick (2022) supply an interpretive caution: performance under irrelevant or misleading instructions need not demonstrate use of intended meaning. Their finding limits inferences from consistency to semantic understanding, but does not establish how models interpret CEFR labels.

Transfer to conditioned writing requires two distinctions. Benchmark accuracy evaluates correctness; the present outcome is a distribution of linguistic features. Accuracy instability need not imply profile instability, and stable profiles need not imply communicative quality. Moreover, similar means can coexist with different dispersion or level-related feature relationships. Location, dispersion, and held-out-wording classification are therefore complementary checks, not interchangeable tests. Agreement supports bounded reproducibility without establishing identical text distributions.

The simulation literature supplies a cautionary precedent for the distributional view. Conditioning a model on persona attributes can reproduce human group averages while distorting the variability and the relationships among variables (Argyle et al., 2023; Bisbee et al., 2024), and replications of human-subject studies with simulated participants succeed unevenly (Aher et al., 2023). By analogy, level conditioning might reproduce a plausible mean profile per level under any reasonable wording while the wordings disagree about dispersion or about the finer feature relationships. This is why location and dispersion are examined separately here, and why RQ3 tests generalization of encodable structure rather than mean agreement alone.

2.3 Task Effects as a Comparison Benchmark

Task effects provide a benchmark for judging wording effects. In human learner corpora, task effects on linguistic complexity and accuracy are substantial and well documented: large-scale analyses of writing in the EFCAMDAT learner corpus show that task identity systematically shifts complexity measures within writers and levels (Alexopoulou et al., 2017). The author's previous generation studies likewise found large task effects in synthetic writing, which is why the present design controls for task in every analysis and reuses a fixed ten-task inventory. Task supplies two useful yardsticks. The variance partition compares the shares attributable to wording, task, and level. In the classifier design, generalization to an unseen wording can be read against generalization to an unseen task, an operation known to be hard. If unseen wordings cost nothing while unseen tasks cost substantially, methodological attention should shift from instruction phrasing to task sampling.

This benchmark also connects the design to the learner-language evidence. Alexopoulou et al. (2017) show why proficiency comparisons must attend to task conditions, while the CEFR complexity studies reviewed above locate linguistic indicators within particular populations and measurement designs. Task variation is therefore not merely residual noise to remove: it helps define the domain over which a level-related pattern can generalize. In the present inventory, task identity bundles topic, genre, and length guidance; comparing its overall effect with wording does not identify the separate contribution of each component. In L2 academic writing contexts, the interaction between prompt design and task structure plays a critical role in shaping output features. Lee et al. (2026) illustrated that curriculum-aligned, task-based AI prompts systematically mediate learner-AI writing interactions, driving a developmental progression from surface-level lexical uptake in early drafting tasks to reflective adaptation at the discourse and argumentation levels.

2.4 Operationalization Invariance and Measurement Validity

Validity concerns the interpretations and uses supported by evidence, rather than an intrinsic property of a score or artifact (Kane, 2013; Messick, 1994). Applied to level-conditioned generation, an argument for using synthetic texts as proficiency-indexed data must specify the inference from an instruction to the resulting linguistic profile and, separately, from that profile to human language proficiency. Dependence on incidental wording would weaken the first inference; independence from wording cannot establish the second. This framing makes paraphrase robustness one component of a validity argument, not a substitute for validation.

The analogy with measurement invariance helps define this component. In psychometrics, invariance concerns whether the relationship between observed measures and a construct is comparable across groups or conditions (Vandenberg & Lance, 2000). Here the instruction is part of the research instrument and its paraphrases are alternative implementations. However, the present analyses do not fit a latent-variable measurement model or test formal measurement invariance. Operationalization invariance is used in a narrower, procedural sense: how far the measured output profiles and their level-related structure change when only the level sentence changes. This distinction prevents a distributional robustness result from being mistaken for evidence that a psychometric scale

has been equated across conditions.

Three claims must therefore be kept separate. Consistency across instruction wordings concerns the reproducibility of measured profiles under paraphrase. Discriminability among requested levels concerns whether those profiles differ sufficiently for level labels to be recovered. Correspondence with human-assessed CEFR proficiency concerns whether the synthetic distinctions represent the intended communicative construct. A model could be consistent and highly discriminable while repeatedly generating a simplified stereotype of proficiency. The present study examines the first two claims, but contains no human benchmark with which to establish the third. This separation follows Kane's (2013) emphasis on specifying and evaluating each inference needed for an intended use, and Messick's (1994) concern with the adequacy of construct representation.

The PSR makes one of these inferential boundaries explicit by expressing wording-related displacement relative to adjacent requested levels. Its denominator is a distance in the generated feature space, not an externally calibrated unit of human proficiency. A small ratio can therefore support procedural consistency without showing that the requested levels have the right substantive meaning. Reporting the raw ratio and its permutation floor separately also distinguishes systematic wording effects from finite-sample scatter; subtracting the floor remains a descriptive aid, not a validated population-distance estimator.

2.5 Implications for Digital L2 Research

Digital L2 research uses generative models as stimulus writers, simulated participants, and data sources. Their calibration belongs within research design: model settings, tasks, instructions, and measures jointly define the procedure. Neither CEFR classification success nor benchmark prompt sensitivity establishes interchangeable proficiency instructions. Reporting exact instructions and testing alternatives makes that assumption inspectable. With machine-generated, machine-measured outcomes, the aim is a reproducible research operation with explicit limits, not a replacement for evidence about learners.

3. Rationale and Research Questions

The literature establishes linguistic correlates of rated proficiency, the feasibility of proficiency-conditioned generation, and the sensitivity of benchmark evaluation to prompts. It does not establish whether equivalent CEFR instructions preserve the measured distributions of generated writing. The empirical gap is consequently about reproducibility under a controlled change in operationalization, not whether prompts can ever matter. The practical question is how far a single instruction can support research claims within a defined task and model domain.

The study crosses four instruction-tuned models, four requested levels (A2 – C1), ten tasks, five wordings, and five replications, yielding 4,000 texts described by twenty prespecified indicators. RQ1 partitions variation to locate wording effects relative to level, model, and task. RQ2 distinguishes centroid displacement from changes in dispersion, using PSR to express displacement relative to adjacent requested levels and a permutation floor to contextualize finite-sample scatter. RQ3 tests whether level-related patterns learned under some wordings generalize to an unseen wording, with an unseen task as a comparison

benchmark. These complementary analyses contribute evidence about procedural consistency and requested-level discriminability; they do not test correspondence with human-assessed proficiency.

Three research questions organize the analyses.

RQ1. How much of the variation in the linguistic characteristics of CEFR-conditioned LLM writing is attributable to the target CEFR level, to the writing task, to the wording of the level instruction, and to the generating model?

RQ2. Does sensitivity to instruction wording differ across CEFR levels and across models, in location (where the average text profile for each level sits) and in dispersion (how variable the outputs are)?

RQ3. Does a CEFR level classifier trained on texts generated under a subset of instruction wordings generalize to texts generated under a semantically equivalent but unseen wording?

4. Method

4.1 Design

The design crosses four generating models, four CEFR levels (A2, B1, B2, C1), ten writing tasks, five instruction wordings, and five replications per cell, for 4,000 requested texts (1,000 per model). Every factor is fully crossed (every combination was run); replications are independent generation requests with no memory of earlier requests, so each cell of the model × level × task × wording design holds five independently generated texts. The two B levels sit between A2 and C1 to give three adjacent level pairs, which the Prompt Sensitivity Ratio uses as its unit of comparison. C2 was excluded, as in the author's previous studies, because task-appropriate C2 writing is weakly distinguishable from C1 on short responses, and A1 because several tasks presuppose content beyond A1 writers.

4.2 Writing Tasks

The ten tasks are those used in the author's previous CEFR generation studies: short argumentative, expository, and narrative prompts drawn from the Write & Improve task inventory (Nicholls et al., 2024), each with its original length guidance (140 to 300 words depending on task). The set includes opinion essays on health, exercise, television, and language learning, expository paragraph tasks on zoos, video games, and teacher continuity, a magazine article on the future of travel, a narrative on a memorable holiday, and a goal-setting report for a new school year. Using the same inventory keeps the task dimension comparable with the earlier findings, ensures the tasks are genuine second-language (L2) writing tasks rather than researcher inventions, and lets task function as a control factor whose large effects are themselves informative as a benchmark. The tasks were originally authored for intermediate learners; following the previous studies, the same task text is used at every requested level, so that level differences in output are attributable to the instruction rather than to task variation.

4.3 Instruction Wordings

Each generation request consisted of a constant frame with one interchangeable sentence. The frame read: "Write a response to the task below as a learner of English. [WORDING] Do not mention CEFR or your proficiency level in your response. Respond naturally as the

learner.", followed by the task text and its length guidance. The five wordings were:

w1. "Write your response at CEFR [level] level."

w2. "Use English appropriate for CEFR level [level]."

w3. "Produce a response corresponding to [level] on the CEFR."

w4. "Write using [level]-level English according to the CEFR."

w5. "Target CEFR [level] proficiency in your response."

The author judged these sentences to preserve the requested target; no independent semantic-equivalence rating was conducted. All five retain the literal CEFR label and level symbol.

The five sentences were constructed to be semantically equivalent requests differing only in surface realization. Each names the CEFR, names the same level symbol, imposes no additional content, and is a fluent single sentence of comparable length; they differ in syntactic frame (imperative with the level as modifier, as prepositional object, or as direct object), in the noun chosen for the output ("response," "English," "proficiency"), and in the verb ("write," "use," "produce," "target"). The set deliberately spans formulations observed in published prompting practice rather than adversarial or degenerate paraphrases; the question is whether ordinary authorial variation matters, not whether sensitivity can be manufactured. Any systematic difference among the five conditions in the generated texts is attributable to wording as such, because nothing else differs.

4.4 Models and Generation

Four instruction-tuned models (models trained to follow written instructions) were selected to span a hosted system from a leading developer, accessed over the internet, and three open-weight models (models whose files can be downloaded and run locally) from different developers and size classes, all usable under reproducible local inference. Claude Sonnet 5.5 (Anthropic, 2026) was invoked through the Claude Code command-line interface in noninteractive mode, with tools, session persistence, and user configuration disabled, a fixed system prompt instructing the model to return only the response, and the vendor's low reasoning-effort setting; each response's metadata was checked to confirm the requested model actually served it. The hosted interface exposes no sampling temperature, top-p, or seed, so exact replay of hosted generations is not claimed; every request and response is retained verbatim.

Qwen3.5-9B (Qwen Team, 2026), Gemma 4 26B-A4B (Google, 2026), and Ministral 3 14B (Mistral AI, 2025) ran locally in quantized (compressed) GGUF form (Q5_K_M, Q4_0, and Q5_K_M respectively, with recorded repository revisions and file hashes) on a llama.cpp server under CUDA 13. Generation used temperature 0.7 and top-p 0.9 (settings that control how much randomness the model uses when choosing each word), a maximum of 1,024 new tokens (word pieces), reasoning modes disabled, and a deterministic per-request seed (a fixed starting number that makes a run repeatable) derived from the request's content hash.

Decoding parameters follow the author's previous studies; sampling (random word selection) rather than greedy decoding (always choosing the most likely word) is required for the dispersion analyses to be meaningful, and the temperature sensitivity of synthetic dispersion was characterized there (see also Holtzman et al., 2020, on decoding effects in

open-ended generation).

Each request was assigned a permanent ID (a hash, or digital fingerprint) computed from its configuration (model, level, task, wording, replication index, prompt text, system prompt, and decoding settings). The original local-model hash also included the machine-specific executable path. That path is redacted in the submission archive, which supplies a checked path-independent fingerprint-to-original-ID-and-seed manifest so all recorded requests can still be reconstructed and generation resumed. New requests use path-independent fingerprints. Requests were issued in a randomized order (fixed by a seed) so that partial runs remain balanced across cells in expectation; raw responses, failures, and manifests are retained.

The same system prompt was used for all four models: "Return only the response to the writing task. Do not add explanations, headings about your answer, or commentary." Generated texts were audited for emptiness, truncation, and level disclosure (any mention of the CEFR or a level symbol in the output), with disclosing or empty responses excluded from analysis by predefined rule.

4.5 Linguistic Indicators

Each text is described by the twenty prespecified indicators used in the author's previous work, computed with spaCy 3.8 (en_core_web_sm; Explosion, n.d.) and wordfreq 3.1.1 (Speer, 2023) counting only alphabetic tokens (words): log word count; mean word length; mean Zipf frequency (a logarithmic scale of how common a word is); rare-word ratio (Zipf < 4); moving-average type-token ratio (MATTR, a vocabulary-variety measure computed over a 50-word window); mean sentence length; mean dependency distance (the average distance between a word and the word it depends on); subordinate-clause ratio; nominal-modifier ratio; noun, verb, adjective, adverb, and pronoun ratios; connective ratio; adjacent-sentence lemma overlap (how many base word forms consecutive sentences share); hedge and booster ratios (words that soften or strengthen a claim); and first- and second-person pronoun ratios.

The set spans the four regions of linguistic description in which CEFR-related variation is typically sought: lexical sophistication and diversity (word length, Zipf frequency, rare words, MATTR), syntactic elaboration (sentence length, dependency distance, subordination, nominal modification), discourse organization (connectives, lemma overlap), and interpersonal stance (hedges, boosters, person reference).

No indicator is model- or wording-aware, and the set was fixed before generation; reusing it unchanged from the previous studies removes a researcher degree of freedom (a choice that could bias results) and keeps effect sizes comparable across the research program. The twenty indicators are a deliberately interpretable probe rather than a maximal representation; the limitation this imposes is taken up in the Discussion and Limitations.

4.6 Analysis

In brief, each text is described by twenty numerical measures, so each text becomes a point defined by twenty numbers. RQ1 asks how much of the spread among these points is associated with level, task, wording, and model (a variance partition). RQ2 asks whether the average position of the texts (location) or their spread (dispersion) changes with the

wording of the instruction. RQ3 asks whether a classifier trained on some wordings still identifies levels in texts written under another wording.

All indicators were *z*-standardized (rescaled to a mean of 0 and a standard deviation of 1) over the analyzed corpus, and Euclidean distance (ordinary straight-line distance) on the standardized vectors defines the feature space (the space in which each text is a point) throughout; with Euclidean distance, the distance-based and multivariate linear decompositions coincide, so PERMANOVA results are exact variance partitions. For RQ1, a permutational multivariate analysis of variance (PERMANOVA; Anderson, 2001) decomposed the feature matrix by task, level, wording, model, and the level × wording, level × model, and model × wording interactions, reporting each term's partial R^2 (the share of total variation attributable to that term once all other terms are accounted for). Significance used Freedman-Lane residual permutation under the reduced model (a procedure that reshuffles what remains after the other terms are removed; Freedman & Lane, 1983) with permutations restricted within tasks (4,999 permutations, making .0002 the smallest attainable *p* value, reported as $p < .001$).

The reported *F* statistic is a multivariate pseudo-*F* aggregated over twenty indicators; its reference distribution is the saved permutation distribution, not the conventional univariate *F* law with the displayed degrees of freedom.

The same decomposition was run within each model (task, level, wording, level × wording). A supplementary model added task × wording and tested it with 999 task-restricted permutations; length sensitivity checks removed log word count and then regressed each remaining indicator on log word count. Three-way and higher interactions were pooled into the residual. The complete analysis protocol and executable scripts are provided in the supplementary data and code archive.

For RQ2, location sensitivity is summarized by the Prompt Sensitivity Ratio (PSR). Within each model, after removing task means, level-by-wording centroids (the average profile of the texts at each level and wording) were computed; for adjacent levels (a, b), PSR = mean within-level pairwise distance among wording centroids, averaged over the two levels of the pair, divided by the mean wording-matched distance between the level centroids. In plain terms, PSR asks how far apart the five wordings' average text profiles are, compared with how far apart two neighboring levels are. PSR near 0 means paraphrase barely moves the estimated profile; PSR near 1 means its observed centroid distance matches an adjacent level change. Approximate confidence intervals use a delete-one-task jackknife *t* method (recomputing the estimate with each task left out in turn) across ten tasks. A 1,000-replicate task bootstrap (recomputing the estimate on tasks resampled with replacement) was also computed; because duplicate-task resampling shifted its distribution upward, its percentile intervals and bias diagnostics are reported in the supplementary output only.

Because averages from a finite sample scatter even when wording has no effect, a permutation noise floor (the PSR expected by chance alone) was computed by shuffling wording labels within model × level × task cells (500 permutations). Observed PSR minus this floor is reported as a descriptive noise-adjusted index, not an unbiased population distance; a permutation *p* value tests whether the observed ratio exceeds the null (smallest attainable value .002).

Dispersion sensitivity (how widely texts vary within each wording) used PERMDISP (a permutation test for differences in spread; Anderson, 2006): within each model, after centering on task × level cells, each text's distance to its wording centroid was tested across wordings by permutation within task × level strata. As a descriptive complement, univariate partial eta squared for wording (the share of one indicator's variation linked to wording) was computed per indicator within each model on the same residuals.

For RQ3, a multinomial logistic regression (a standard classification model) on the twenty indicators (standardized within training folds) predicted the four-way level label under three partitions of each model's texts: stratified random five-fold (five random splits that keep the levels balanced); leave-one-wording-out (train on four wordings, test on the fifth, rotating); and leave-one-task-out as a severity benchmark (a harder test, because the classifier faces new topics). Accuracy, macro-F1 (a score that weights all four levels equally), and adjacent accuracy (within one level) are reported from pooled held-out predictions; because every text receives a prediction under each scheme, the random-minus-wording accuracy difference is paired at the text level and carries a 95% task-cluster bootstrap interval (1,000 replicates; tasks were resampled as whole units).

The analysis seed, permutation counts, and software versions are fixed in the supplementary archive; six scientific unit tests verify permutation calibration under null and non-null synthetic data, PSR behavior on constructed geometries, partition integrity, and dispersion detection.

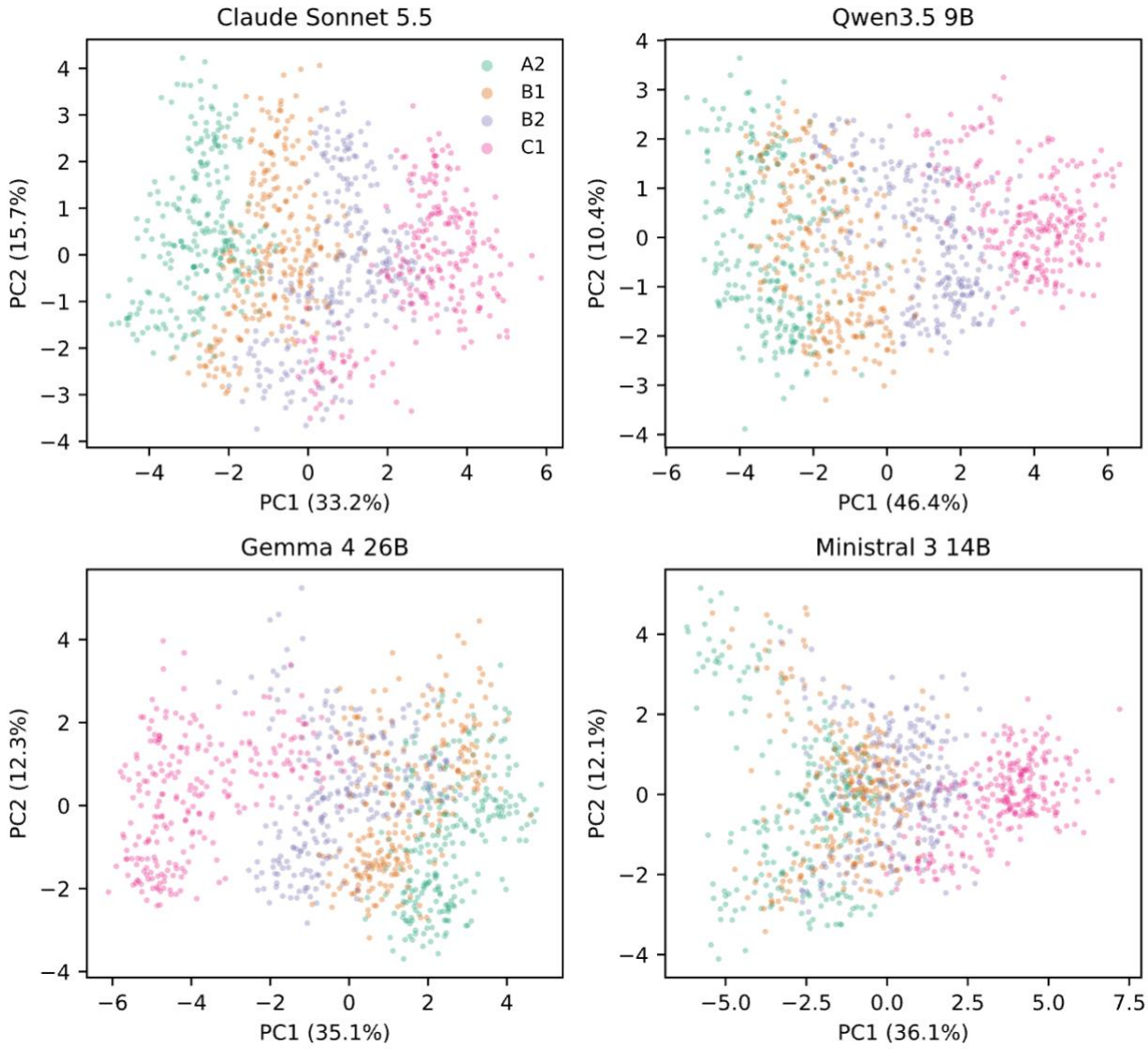
Three methodological choices deserve brief justification. PERMANOVA was preferred over parametric MANOVA because the indicator distributions are bounded ratios and counts whose multivariate normality cannot be assumed; Freedman-Lane permutation conditions each term on the others while respecting task blocking. The task-cluster bootstrap for classifier intervals and the jackknife for PSR treat the ten tasks as sampled units; with so few clusters, uncertainty intervals are approximate and coverage is unverified. The within-cell permutation floor for PSR estimates the ratio expected from finite-sample centroid scatter under no wording effect. Subtracting this floor aids description but does not recover an unbiased population displacement.

5. Results

How to read the statistics. R^2 is the proportion of the total variation in the twenty indicators that is associated with a factor (for example, .278 means 27.8%). F is a test statistic that grows as a factor's effect grows relative to the variation left unexplained, and p is the probability of a result at least this extreme if the factor had no effect; p values below .05 are conventionally called statistically significant. *SD* is the standard deviation. Because a statistically significant effect can still be very small, the effect sizes (R^2, the Prompt Sensitivity Ratio, and accuracy differences) matter more here than the p values.

5.1 Generation Audit and Descriptives

All 4,000 requests completed on the first pass with no failures, refusals, empty responses, or truncations: every local generation terminated at a natural stopping point and every hosted generation ended its turn normally. No generated text disclosed a CEFR level despite every instruction naming one, so no exclusions were required, and every one of the 800 model × level × task × wording cells holds exactly five texts. Mean response length ranged from 196 words (*SD* = 31) for Gemma 4 26B-A4B to 246 words (*SD* = 47) for Ministral 3 14B, with Claude Sonnet 5.5 at 224 (*SD* = 31) and Qwen3.5-9B at 231 (*SD* = 33). Length rose modestly with requested level, from 210 words (*SD* = 34) at A2 to 240 (*SD* = 48) at C1 pooled over models, almost entirely because of Ministral, whose means climbed from 215 at A2 to 286 at C1 while the other three models stayed within a 15-word range across levels. Figure 1 plots the texts in a two-dimensional summary of the twenty indicators (principal component analysis, PCA, fitted separately within each model); the variance explained by each plotted component appears on its axis and is tabulated in Table S5 of the supplementary analysis. The first component orders requested levels A2 through C1, with overlap between adjacent levels.



Note. PCA was fitted separately for each model on the standardized indicators. PC1 and PC2 (the first two principal components) variance shares are printed on each panel's axes. Points are individual texts pooled over tasks, wordings, and replications.

Figure 1. Indicator Space by Requested Level

Using the same alphabetic-token count as the feature table, 1,781 of 4,000 responses (44.5%) exceeded their task's original maximum length, and one fell below its minimum (see Table S6 in the supplementary analysis). Above-maximum rates differed by model: 1.0% for Gemma, 51.8% for Claude, 62.4% for Ministral, and 62.9% for Qwen. These responses were retained because length guidance was identical across the five wording conditions; exceeding it is an instruction-following outcome, not an exclusion criterion. The above-maximum share ranged from 43.5% to 45.8% across wordings.

The texts themselves read as the conditioning intends, independently of which wording requested them. A representative A2 response opens "I think video games are very popular with children because they are fun and exciting. Many kids love to play them after school." (Qwen3.5-9B, task 3, wording w1), in short declaratives built from high-frequency vocabulary; a C1 response to a different task opens "Undoubtedly, regular physical activity is a cornerstone of a healthy lifestyle, offering profound benefits ranging from

cardiovascular improvement to mental clarity." (Qwen3.5-9B, task 5, wording w3), with fronted stance marking, nominalization, and low-frequency lexis. Informal inspection across wordings found no output whose register, length, or apparent level would allow a reader to guess which of the five instruction sentences produced it, an impression the formal analyses below quantify.

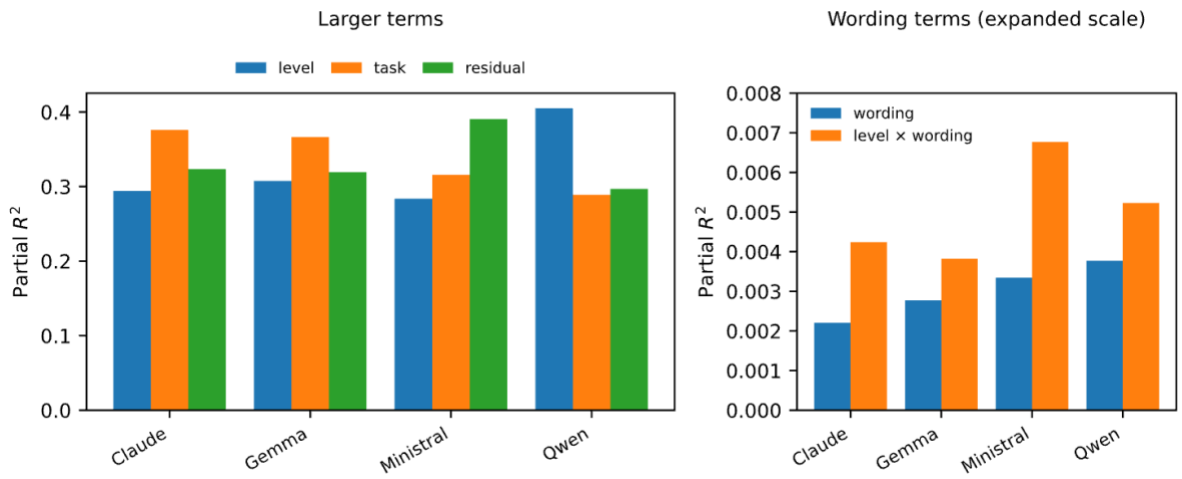
5.2 RQ1: Variance Partition

Table 1 and Figure 2 report the distance-based decomposition. In the overall partition, the requested level accounted for $R^2 = .278$ of total indicator variance ($F(3, 3947) = 1030.4$, $p < .001$), essentially tied with the task ($R^2 = .280$, $F(9, 3947) = 346.6$) and far ahead of the generating model ($R^2 = .058$, $F(3, 3947) = 213.9$, $p < .001$) and the level × model interaction ($R^2 = .026$, $F(9, 3947) = 31.5$, $p < .001$), the latter two confirming that models differ in both their baseline styles and the steepness of their level realization. Instruction wording accounted for $R^2 = .0012$ ($F(4, 3947) = 3.4$, $p < .001$), the level × wording interaction for $R^2 = .0014$ ($F(12, 3947) = 1.3$, $p = .006$), and the model × wording interaction for $R^2 = .0017$ ($F(12, 3947) = 1.5$, $p < .001$). The permutation audit reproduced all three p values: the 99th null percentiles (the values exceeded by only 1% of permuted results) were 1.56, 1.30, and 1.30, respectively, versus observed F values of 3.38, 1.34, and 1.54 (see Table S1 and Figure S1 in the supplementary analysis for the empirical null distributions). Wording effects are detectable at this sample size, but their variance shares remain small.

Table 1. Distance-Based Variance Partition of the Twenty Indicators

Term	df	R^2	F	p
Task	9	.280	346.6	
Level	3	.278	1030.4	< .001
Wording	4	.001	3.4	< .001
Level × Wording	12	.001	1.3	.006
Model	3	.058	213.9	< .001
Level × Model	9	.025	31.5	< .001
Model × Wording	12	.002	1.5	< .001
Residual	3947	.355		
Total	3999	1.000		

Note. Partial R^2 is each term's share of total sum of squares on standardized features under Euclidean distance. p values are from 4,999 Freedman–Lane permutations restricted within tasks; the task block is not tested. Per-model partitions appear in the supplementary tables.



Note. Partial R^2 by term and model from the within-model variance partitions. The right panel expands the scale for wording and level × wording.

Figure 2. Variance Partition by Model

The within-model partitions sharpen the picture. Task shares ranged from R^2 = .289 (Qwen3.5-9B) to .376 (Claude Sonnet 5.5), and level shares from .284 (Ministral 3 14B) through .294 (Claude Sonnet 5.5) and .308 (Gemma 4 26B-A4B) to .405 (Qwen3.5-9B). Against these, the wording share ranged from R^2 = .0022 (Claude Sonnet 5.5, p = .006) to .0038 (Qwen3.5-9B, $p < .001$), with Gemma at .0028 and Ministral at .0033 (both $p < .001$). The level × wording interaction reached significance only for Ministral 3 14B (R^2 = .0068, p = .001) and Qwen3.5-9B (R^2 = .0052, p = .001), not for Claude Sonnet 5.5 (R^2 = .0042, p = .301) or Gemma 4 26B-A4B (R^2 = .0038, p = .574). These results concern separation among requested levels, not fidelity to human CEFR proficiency.

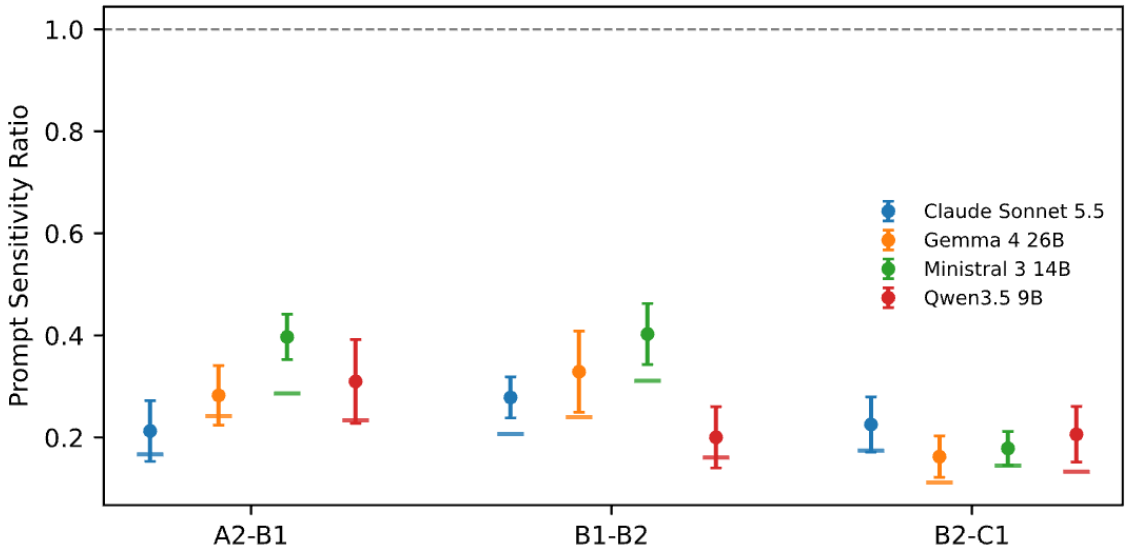
5.3 RQ2: Location and Dispersion Sensitivity

Table 2 and Figure 3 report the Prompt Sensitivity Ratio. PSR values near 0 mean that rewording barely changes the texts; values near 1 mean that rewording changes them as much as moving up one CEFR level. Observed model means were .24 (Claude Sonnet 5.5 and Qwen3.5-9B), .26 (Gemma 4 26B-A4B), and .33 (Ministral 3 14B). Permutation floors (the PSR expected by chance alone) were .18, .18, .20, and .25 respectively, leaving observed-minus-floor excesses of .06, .06, .06, and .08 (all model-level permutation p = .002). These are three distinct quantities: the raw ratio, its null expectation, and their descriptive difference. The difference is not an unbiased estimate of underlying displacement. At the pair level, Ministral reached PSR = .40 at A2–B1 (excess .11) and .40 at B1–B2 (excess .09); its B2–C1 excess was .03. The ten-task bootstrap means were .048 to .056 above the observed model means, and several percentile lower bounds exceeded their point estimates. Duplicated tasks increase variability of estimated centroid differences, a plausible source of upward shift. We therefore display approximate delete-one-task jackknife t intervals, retain both interval types in the analysis outputs, and make no coverage claim for either with only ten tasks (see Table S2 in the supplementary analysis for both interval summaries and the noise floor).

Table 2. Prompt Sensitivity Ratio by Model and Adjacent Level Pair

Model	A2–B1	B1–B2	B2–C1	Mean	Noise floor	Excess
Claude Sonnet 5.5	0.21 [0.15, 0.27]	0.28 [0.24, 0.32]	0.23 [0.17, 0.28]	0.24 [0.20, 0.28]	0.18	0.06
Gemma 4 26B-A4B	0.28 [0.22, 0.34]	0.33 [0.25, 0.41]	0.16 [0.12, 0.20]	0.26 [0.21, 0.31]	0.20	0.06
Ministral 3 14B	0.40 [0.35, 0.44]	0.40 [0.34, 0.46]	0.18 [0.15, 0.21]	0.33 [0.30, 0.36]	0.25	0.08
Qwen3.5 9B	0.31 [0.23, 0.39]	0.20 [0.14, 0.26]	0.21 [0.15, 0.26]	0.24 [0.20, 0.28]	0.18	0.06

Note. PSR is the mean distance among wording centroids within a level divided by the wording-matched adjacent-level distance, after removing task means. Brackets give approximate 95% delete-one-task jackknife *t* intervals (10 tasks). The task bootstrap distribution was shifted upward by duplicate-task resampling; its diagnostics are supplied with the analysis. The noise floor is the mean PSR under 500 within-cell wording-label permutations. Excess is observed mean minus floor, a descriptive index rather than an unbiased population distance (all model-level permutation p = .002).



Note. Prompt Sensitivity Ratio by model and adjacent level pair with approximate 95% delete-one-task jackknife *t* intervals. Horizontal dashes mark each estimate's permutation noise floor; the dashed line marks parity.

Figure 3. Prompt Sensitivity Ratio

The dispersion test (PERMDISP) detected no statistically significant wording differences in dispersion (Claude Sonnet 5.5 $F = 0.93$, $p = .296$; Gemma $F = 0.45$, $p = .708$; Ministral $F = 0.87$, $p = .360$; Qwen $F = 1.57$, $p = .105$). This does not establish equivalence. Average distances from each text to its wording group's centroid differed across wordings by at most 0.12 standardized units. Claude Sonnet 5.5's within-cell dispersion rose from A2 (1.75) to C1 (2.07), whereas Ministral's decreased from 2.55 to 2.37.

5.4 Where Wording Effects Concentrate

The univariate decomposition locates the small wording effects in interpretable places. The largest single-indicator effect anywhere in the study was Gemma's mean Zipf word frequency (partial eta squared = .072), with rare-word ratio second in the same model (.042): relative to the other wordings, the formulation "Produce a response corresponding to [level] on the CEFR" (w3) shifted Gemma's vocabulary toward more frequent words by about 0.19 standardized units and reduced rare words correspondingly. The same w3 direction appears, attenuated, in Claude Sonnet 5.5's mean Zipf frequency (eta squared = .031). Qwen's largest effect was response length (log words, .048), with w1 and w5 producing slightly shorter texts than the middle wordings. Ministral's largest effect was mean sentence length (.035), a syntactic measure: "Target CEFR [level] proficiency" (w5) lengthened sentences by 0.28 standardized units relative to w4. No indicator in any model exceeded eta squared = .08. The largest observed shifts thus involved vocabulary frequency, text length, and sentence length; the latter shows that syntax was not entirely unaffected.

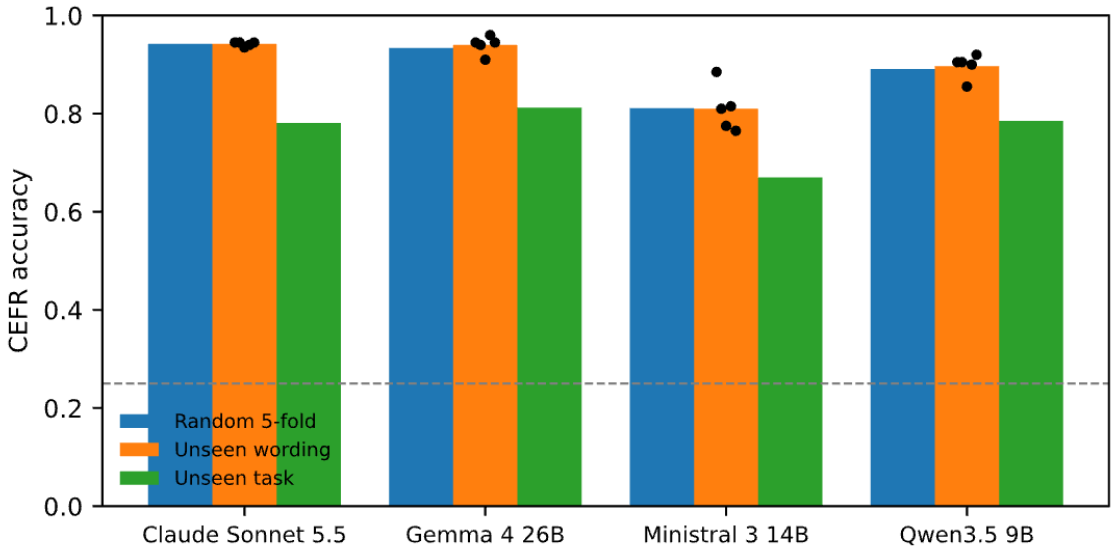
5.5 RQ3: Generalization to an Unseen Wording

Table 3 and Figure 4 report the classifier results. Under random five-fold cross-validation (training on four fifths of the texts and testing on the remaining fifth, five times), the twenty indicators recovered the requested level at 94.2% (Claude Sonnet 5.5), 93.4% (Gemma 4 26B-A4B), 89.1% (Qwen3.5-9B), and 81.1% (Ministral 3 14B) against a 25% chance rate, with macro-F1 within a tenth of a point of accuracy throughout. Evaluating the same texts under leave-one-wording-out partitions changed these accuracies by at most six tenths of a point: pooled differences were .000 (95% confidence interval, CI, [-.008, .008]) for Claude Sonnet 5.5, -.006 ([-.017, .004]) for Gemma, .001 ([-.004, .006]) for Ministral, and -.006 ([-.012, -.001]) for Qwen, whose held-out wordings were classified very slightly better than random folds. Adjacent accuracy under unseen wordings was at least .999 in every model: when a classifier facing an unfamiliar wording errs at all, it errs by one level. Individual held-out-wording folds were tightly grouped for Claude Sonnet 5.5 (.935 to .945) and Gemma (.910 to .960), looser for Qwen (.855 to .920), and widest for Ministral (.765 to .885), consistent with that model's larger PSR excess and significant level × wording interaction.

Table 3. CEFR Classification Accuracy Under Three Cross-Validation Partitions

Model	Random	Unseen wording	Unseen task	Δ Random − wording [95% CI]	Adjacent (wording)
Claude Sonnet 5.5	.942	.942	.781	0.000 [-0.008, 0.008]	1.000
Gemma 4 26B-A4B	.934	.940	.812	-0.006 [-0.017, 0.004]	1.000
Ministral 3 14B	.811	.810	.670	0.001 [-0.004, 0.006]	.999
Qwen3.5 9B	.891	.897	.785	-0.006 [-0.012, -0.001]	.999

Note. Multinomial logistic regression on the twenty indicators, within model. Random = stratified five-fold; unseen wording = leave-one-wording-out; unseen task = leave-one-task-out. Chance is .25. Δ is the pooled accuracy difference with a 95% task-cluster bootstrap interval. Adjacent = within one level under the unseen-wording partition.



Note. CEFR classification accuracy per model under random five-fold, leave-one-wording-out, and leave-one-task-out partitions. Black points show the five individual held-out-wording folds; the dashed line marks chance.

Figure 4. Classifier Generalization Across Instruction Wordings

The task benchmark puts these numbers in perspective. Leave-one-task-out evaluation cost 10.6 to 16.1 accuracy points (Claude Sonnet 5.5: .942 to .781; Gemma: .934 to .812; Qwen: .891 to .785; Ministral: .811 to .670), twenty to one hundred times the wording cost, while adjacent accuracy stayed at .984 or above. What looks like CEFR encoding in these systems transfers essentially without loss to an instruction wording never seen in training, while transferring to an unseen task is measurably and consistently harder.

5.6 Robustness

The supplementary interaction accounted for $R^2 = .0026$ ($F(36, 3911) = .79$, task-restricted permutation $p = .998$). Within individual tasks, wording shares ranged from .0036 to .0068. Removing log word count left the task share at .272 and the wording share at .0013; statistically removing log word count from each remaining indicator left .266 and .0013 (see Table S3 in the supplementary analysis). Without log word count, unseen-task classifier accuracy was .662 to .813, whereas unseen-wording accuracy was .809 to .944. With fold-specific length adjustment, the ranges were .531 to .768 and .612 to .915, respectively (see Table S4 in the supplementary analysis for model-specific accuracies). Thus, length explains part of the task contrast, especially for the hosted model, but does not erase it.

The permutation procedures passed six software tests run on synthetic data. The reported $p < .001$ values are the smallest values that 4,999 permutations can produce, and the PSR model-level p values are the smallest that 500 permutations can produce. With only ten tasks, the PSR intervals remain approximate; neither the bootstrap percentiles nor the jackknife intervals have demonstrated nominal coverage.

6. Discussion

6.1 Summary of Findings

Across four models and 4,000 texts, the requested level and task each accounted for roughly 28% of indicator variance, while wording accounted for about 0.1%. No statistically significant wording differences in dispersion were detected. A nonsignificant PERMDISP test does not demonstrate equivalent dispersion; no equivalence margin was tested. A level classifier trained on four wordings changed by at most .006 accuracy on the fifth; unseen tasks cost .106 to .161. Wording effects were detectable in every model and level-dependent in two. The narrow conclusion is substantial stability across these five English paraphrases, alongside model and task differences that remain worth checking.

6.2 Interpreting Stability Within the Tested Conditions

The results extend Mizrahi et al.'s (2024) evaluation question to a different operation: relatively stable linguistic profiles and held-out-wording classification contrast with paraphrase-sensitive benchmark performance. Differences in tasks, models, manipulations, and outcomes preclude concluding that generation is intrinsically more robust. Example-order and formatting interventions (Lu et al., 2022; Sclar et al., 2024) were not tested here. Text-level aggregation may attenuate wording differences or conceal unmeasured ones; neither explanation was demonstrated.

All five wordings retained CEFR and the proficiency symbol, potentially providing a strong shared conditioning cue. This tests variation around a common label, not broad semantic invariance. Behavioral consistency does not establish interpretation of intended meaning (Webson & Pavlick, 2022). Without label-removal, descriptor-only, or translated conditions, shared-token conditioning cannot be distinguished from instruction-tuning robustness or other explanations. It remains a hypothesis and boundary condition, not a demonstrated mechanism.

Although wording effects were relatively small in the present design, establishing controlled prompting frameworks remains essential for applied pipelines. When generative AI is used for academic text revision or L2 writing support, uncurated heavy prompting risks displacing authorial voice and introducing uncalibrated rare vocabulary (Watson Todd, 2026). Conversely, embedding structured prompt decision-points into task-based frameworks encourages evaluative engagement and metacognitive reflection in L2 writing tasks (Lee et al., 2026).

6.3 Implications for Research Practice

Task identity explained as much indicator variance as requested level, and unseen tasks cost classifier accuracy 11–16 points. Length adjustment reduced some contrasts but left task effects dominant. This connects conceptually with learner-corpus task effects (Alexopoulou et al., 2017) and context-dependent CEFR complexity (Khushik & Huhta, 2020, 2022; Yamashita, 2025), without establishing common magnitudes or mechanisms in synthetic and human writing. Tasks bundle topic, genre, and length guidance; their causal contributions were not isolated. Generalization therefore requires task coverage as well as alternative phrasings.

Model differences also matter. Qwen showed the strongest separation among requested levels (R^2 = .405), while Ministral showed the weakest (R^2 = .284) and the largest PSR. These are properties of prompted-level discriminability, not measures of CEFR fidelity.

Reporting should include the complete instruction verbatim, the number of formulations, and, where feasible, PSR or a wording variance share from a factorial probe. Sensitivity checks should reflect the intended use and model, task, and wording domain;

these results establish no universal minimum sample size. Such reporting distinguishes demonstrated robustness from an unchecked assumption.

6.4 Implications for Applied Systems

For synthetic corpora, materials pipelines, and assessment prototypes, wording was a minor factor within the tested conditions. Training on four wordings supported classification under the fifth, within .006 accuracy of the random-split baseline. Diversification remains prudent where wording effects are larger or level-dependent. Small effects concentrated in vocabulary frequency and length can still matter for tightly controlled readability: a 0.19 standardized-unit vocabulary shift may be consequential when tuning vocabulary load to a fraction of a requested level. Such applications should fix wording or sample several formulations.

6.5 Stability Is Not Validity

Consistency across wordings and discriminability among requested levels do not establish correspondence with human CEFR proficiency. Linguistic correlates of rated learner levels cannot automatically validate synthetic distinctions (Wisniewski, 2017). Prior comparisons identify synthetic–human differences in register and stance (Carlo & Takeuchi, 2026b, 2026c), but this experiment includes no human benchmark. In Kane's (2013) terms, the evidence supports a reproducible measured generation target, not the further inference to human proficiency. Consistency cannot compensate for incomplete construct representation (Messick, 1994); that inference requires independent human ratings, comparable learner texts, and broader communicative criteria.

6.6 Future Directions

Three extensions follow. The first replaces the level symbol with CEFR descriptors, examples, or translated instructions to test whether the shared token anchors stability. The second varies prompt format or system-versus-user placement. The third calibrates synthetic level distances against human-assessed writing, so the denominator of PSR has an external proficiency anchor rather than only model-internal spacing.

6.7 A Reusable Instrument

The Prompt Sensitivity Ratio can be adapted to other graded generation targets, such as formality or readability, when an adjacent target step is meaningful. Raw PSR, a within-cell permutation floor, and their descriptive difference should be reported separately. Their interpretation depends on the feature space, number of replications, and validity of the target step; they should not be treated as directly comparable population displacements across studies.

7. Limitations and Conclusion

Six constraints bound the conclusions. First, five fluent English wordings that all name CEFR and the level symbol sample a narrow paraphrase space. Second, each model ran under one decoding configuration; the hosted model exposes no seed, temperature, or top-p, so exact replay is unavailable. Third, twenty indicators may miss wording signatures present elsewhere in the text, and classifier accuracy nears a ceiling in two models. Fourth, four models at one point in time do not establish a general property of language models. Fifth, the study includes no human benchmark and cannot validate the prompted levels as CEFR proficiency. Sixth, tasks bundle topic, genre, and prescribed length; although length-adjusted checks preserve a large task effect, they do not isolate these components.

Within the tested models, tasks, indicators, and five label-preserving English paraphrases, CEFR-conditioned generation was a reproducible research operation on the measured outcomes, with wording effects small relative to task variation. Unseen wording changed classifier accuracy by at most .006, whereas an unseen task cost .106 to .161; task effects remained substantial after length checks. The raw PSR and its permutation floor describe this sensitivity, but their difference is not a population displacement. Broader paraphrases, additional tasks, and human-assessed writing are needed to test generality and CEFR validity.

Data Availability

The supplementary analysis, supplied at https://osf.io/bdce8/overview?view_only=1f193c8b2b324eb1a8d6bbafdcae0eb9, contains Tables S1 – S6 and Figure S1. The separate supplement_anonymized.zip archive contains the 4,000 generated texts, feature table, portable request manifest, complete analysis protocol, scripts, tests, saved permutation-null distributions, and analysis outputs. The archive has not yet been deposited at a public address.

Declaration of GenAI Use in the Writing Process

Claude Sonnet 5.5, Qwen3.5-9B, Gemma 4 26B-A4B, and Ministral 3 14B generated the experimental texts described in Method. Claude Sonnet 5.5 assisted with legibility, conciseness, and editorial corrections for the manuscript.

Ethical Statement

Not Applicable: Ethical approval is not applicable for this article.

Funding

This research received no external funding.

Conflict of Interest

The author declares no conflicts of interest regarding this manuscript.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 337–371). PMLR. https://proceedings.mlr.press/v202/aher23a.html

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning, 67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology, 26*(1), 32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics, 62*(1), 245–253. https://doi.org/10.1111/j.1541-0420.2005.00440.x

Anthropic. (2026). *Claude Sonnet 5.5* [Large language model]. https://www.anthropic.com

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Carlo, M., & Takeuchi, O. (2026a). Do large language models encode second-language writing proficiency? A CALF-based perspective. *Digital Studies in Language and Literature, 3*(1), 97–142. https://doi.org/10.1515/dsll-2025-0019

Carlo, M., & Takeuchi, O. (2026b). Form without function? Register sensitivity in CEFR-conditioned LLM and second-language learner writing. *International Journal of Applied Linguistics*. Advance online publication. https://doi.org/10.1111/ijal.70284

Carlo, M., & Takeuchi, O. (2026c). Proficiency without interaction: Comparing stance and engagement in LLM-generated and human CEFR writing. *International Journal of Applied Linguistics*. Advance online publication. https://doi.org/10.1111/ijal.70353

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://rm.coe.int/16809ea0d4

Explosion. (n.d.). *English trained pipelines* [Computer software documentation]. https://spacy.io/models/en

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics, 1*(4), 292–298. https://doi.org/10.1080/07350015.1983.10509354

Google. (2026). *Gemma 4 26B-A4B instruction-tuned QAT Q4_0 GGUF* [Large language model]. Hugging Face. https://huggingface.co/google/gemma-4-26B-A4B-it-qat-q4_0-gguf

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal, 1*(1), Article e5. https://doi.org/10.1017/S2041536210000103

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rygGQyrFvH

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kerz, E., Wiechmann, D., Qiao, Y., Tseng, E., & Ströbel, M. (2021). Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 199–209). Association for Computational Linguistics. https://aclanthology.org/2021.bea-1.21/

Khushik, G. A., & Huhta, A. (2020). Investigating syntactic complexity in EFL learners' writing across Common European Framework of Reference levels A1, A2, and B1. *Applied Linguistics, 41*(4), 506–532. https://doi.org/10.1093/applin/amy064

Khushik, G. A., & Huhta, A. (2022). Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1–B2. *European Journal of Applied Linguistics, 10*(1), 142–184. https://doi.org/10.1515/eujal-2021-0011

Lee, Y.-J., Parrila, R., & Colenbrander, D. (2026). Adopt, adapt, or reject? Analysing student-AI interaction in university curriculum-aligned writing tasks. *Applied Language Sciences, 2,* 260102. https://doi.org/10.65553/ALS.260102

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8086–8098). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.556

Malik, A., Mayhew, S., Piech, C., & Bicknell, K. (2024). From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15670–15693). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.926

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (Research Report No. RR-94-45). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

Mistral AI. (2025). *Ministral 3 14B Instruct 2512 GGUF* [Large language model]. Hugging Face. https://huggingface.co/mistralai/Ministral-3-14B-Instruct-2512-GGUF

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics, 12*, 933–949. https://doi.org/10.1162/tacl_a_00681

Nicholls, D., Caines, A., & Buttery, P. (2024). *The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English*. Cambridge University Press & Assessment. https://doi.org/10.17863/CAM.112997

Qwen Team. (2026). *Qwen3.5-9B* [Large language model]. Hugging Face. https://huggingface.co/Qwen/Qwen3.5-9B

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*. https://openreview.net/forum?id=RIu5lyNXjT

Speer, R. (2023). *wordfreq* (Version 3.1.1) [Computer software]. Python Package Index. https://pypi.org/project/wordfreq/3.1.1/

Vajjala, S., & Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 147–153). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0515

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Watson Todd, R. (2026). Using generative AI to revise articles: How much is too much? *Applied Language Sciences, 2*, 260108. https://doi.org/10.65553/ALS.260108

Webson, A., & Pavlick, E. (2022). Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2300–2344). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.167

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning, 67*(S1), 232–253. https://doi.org/10.1111/lang.12223

Yamashita, T. (2025). Examining English language learners' longitudinal development of syntactic complexity across five CEFR levels with a robust measurement design: A mixed-methods approach. *Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amaf074

Author biodata

Michele Carlo is a doctoral researcher in Applied Linguistics at Kansai University, Japan. His research focuses on second-language writing, generative artificial intelligence, automated language assessment, and the validity and transparency of large language model‑based evaluation. He is particularly interested in how linguistic form, function, and proficiency are represented in AI-generated and human learner writing, and in developing reproducible, privacy-conscious research workflows using open-weight models. His recent work has examined LLM scoring, stance and engagement, CEFR-related linguistic variation, and methodological transparency in applied linguistics. He also teaches English and academic communication at several universities in Japan.
E-mail: m.carlo@me.com
ORCID: https://orcid.org/0009-0007-9011-6531