Appendix

A Datasheet for Datasets

The following section is answers to questions listed in datasheets for datasets.

A.1 Motivation

- Question: **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

  Answer: To evaluate the linguistic robustness of language models across diverse English varieties by transforming Standard American English (SAE) datasets.
- Question: **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

  Answer: The authors of this paper.
- Question: **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

  Answer: This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)).

A.2 Composition

- Question: **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

  Answer: QA datasets (sentences) transformed into various English varieties.
- Question: **How many instances are there in total (of each type, if appropriate)?**

  Answer: There are about 952K instances in total.
- Question: **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

  Answer: The dataset contains all instances from the existing benchmark datasets.
- Question: **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

  Answer: Each instance consists of the transformed text, answer choices, and label.
- Question: **Is there a label or target associated with each instance?** If so, please provide a description.

  Answer: Yes, each label comes from the original QA datasets.
- Question: **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

  Answer: No, there is no information missing from individual instances.
- Question: **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

  Answer: No.
- Question: **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

Answer: This dataset is for testing only.

- Question: **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

  Answer: No, we have verified that there are no errors in the datasets.

- Question: **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

  Answer: Our dataset is self-contained.

- Question: **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** If so, please provide a description.

  Answer: No.

- Question: **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

  Answer: No.

A.3 Preprocessing/cleaning/labeling

- Question: **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remaining questions in this section.

  Answer:

- Question: **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

  Answer: No.

- Question: **Is the software that was used to preprocess/clean/label the data available?** If so, please provide a link or other access point.

  Answer:

  - Google Sheets: `https://docs.google.com/spreadsheets/`
  - Python: `https://www.python.org/`

A.4 Uses

- Question: **Has the dataset been used for any tasks already?** If so, please provide a description.

  Answer: No.

- Question: **Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

  Answer: No.

- Question: **What (other) tasks could the dataset be used for?**

  Answer: N/A

- Question: **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

Answer: N/A

- Question: **Are there tasks for which the dataset should not be used?** If so, please provide a description.

  Answer: N/A

A.5 Distribution

- Question: **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

  Answer: Yes, the dataset will be made publicly accessible through Hugging Face.
- Question: **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?

  Answer: The datasets will be distributed on Hugging Face with public access.
- Question: **When will the dataset be distributed?**

  Answer: The dataset is publicly available on Hugging Face since May 12, 2025.
- Question: **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

  Answer: The datasets are distributed under the CC BY-SA 4.0 license.
- Question: **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

  Answer: No.
- Question: **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

  Answer: No.

A.6 Maintenance

- Question: **Who will be supporting/hosting/maintaining the dataset?**

  Answer: The dataset is hosted on Hugging Face.
- Question: **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

  Answer: Contact the authors of this paper via email.
- Question: **Is there an erratum?** If so, please provide a link or other access point.

  Answer: No.
- Question: **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

  Answer: The datasets will be updated if necessary.
- Question: **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

  Answer: The dataset does not relate with people.

- Question: **Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

  Answer: Yes.

- Question: **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

  Answer: No, our datasets are freely available for others to use.

B Experiment Setting

B.1 Computer Resources

Experiments were conducted using four NVIDIA RTX A6000 GPUs and two NVIDIA A100-SXM4-80GB GPUs. Our implementation is built on vLLM (v0.5.5), PyTorch (v2.4.0), Hugging Face Transformers (v4.47.0), and Datasets (v3.1.0). On average, each dataset required approximately 10 hours for transformation.

C Dataset Construction

C.1 English Dialects

C.1.1 Electronic World Atlas of Varieties of English (eWAVE)

The Electronic World Atlas of Varieties of English (eWAVE) [34] is a curated database documenting 235 linguistic features across 77 English varieties. Developed by 84 professional linguists and grounded in 175 peer-reviewed sources, eWAVE provides a structured taxonomy of features spanning 12 grammatical categories: Pronouns, Noun Phrase, Tense and Aspect, Modal Verbs, Verb Morphology, Negation, Agreement, Relativization, Complementation, Adverbial Subordination, Adverbs and Prepositions, and Discourse and Word Order. Each feature is accompanied by illustrative examples. Varieties are annotated with six levels of feature prevalence: (i) feature is pervasive or obligatory, (ii) feature is neither pervasive nor extremely rare, (iii) feature exists, but is extremely rare, (iv) attested absence of feature, (v) feature is not applicable (given the structural make-up of the variety/P/C), and (vi) no information on feature is available.

C.1.2 Dialect Selection

We first mapped the presence strength of each feature per dialect to one of four discrete levels.

- feature is pervasive or obligatory: 1.0
- feature is neither pervasive nor extremely rare: 0.5
- feature exists, but is extremely rare: 0.25
- attested absence of feature, feature is not applicable, no information on feature is available: 0

We then applied Singular Value Decomposition (SVD) for dimensionality reduction, retaining 90% of the variance. Using the reduced feature representations for each dialect, we performed K-Nearest Neighbors (KNN) clustering with the number of clusters set to 5. The choice of 5 clusters was informed by both the Elbow Method and Silhouette Scores, which indicated that 5 was the most optimal number of clusters. Then we selected clusters with famous English dialects such as African American Vernacular English and Welsh English. The final 18 dialects and their abbreviations are as follows: African American Vernacular English (AAVE), Irish English (IrE), Australian English (AuE), Bahamian English (BahE), East Anglian English (EAngE), Appalachian English (AppE), English dialects in the Southeast of England (SE-Eng), Australian Vernacular English (AuE-V), English dialects in the North of England (NE-Eng), English dialects in the Southwest of England (SW-Eng), Manx English (Manx), New Zealand English (NZE), Newfoundland English (NfE), Ozark

English (OzE), Scottish English (ScE), Southeast American enclave dialects (SE-AmE), Tristan da Cunha English (TdCE), Welsh English (WaE).

C.2 ESL English-L1

C.2.1 Number of Samples in Compiled Dataset.

Table 6 shows the number of samples per L1 and per CEFR level collected from three learner corpora: CLC-FCE [68], ICLE [22], and EFCamDat [18].

Table 6: Number of samples collected from CLC-FCE, ICLE, and EFCamDat.

	CLC-FCE		ICLE		EFCamDat		Total	
	A	B	A	B	A	B	A	B
Arabic	0	0	0	0	24,155	4,857	24,155	4,857
Chinese-Mandarin	9	107	1	45	106,654	22,289	106,664	22,441
French	2	245	0	0	22,244	9,646	22,246	9,891
German	2	120	3	42	25,040	14,501	25,045	14,663
Italian	2	121	1	8	22,787	11,672	22,790	11,801
Japanese	6	134	10	171	11,653	5,081	11,669	5,386
Portuguese	1	114	1	43	248,200	61,751	248,202	61,908
Russian	10	134	0	12	35,081	13,287	35,091	13,433
Spanish	16	351	6	47	52,786	11,456	52,808	11,854
Turkish	8	126	0	61	7,899	2,237	7,907	2,424

C.2.2 CEFR Pseudo-Label Generation

The CLC-FCE and ICLE datasets do not include annotated CEFR levels. To address this, we employed `gpt-4o-mini-2024-07-18` to generate pseudo-CEFR labels. The prompt used for label generation is provided in Table 7.

Table 7: Prompt used for pseudo CEFR label generation.

System:
You are a linguistic expert.

User:
Classify the given sentence among three CEFR levels (A, B, C). Respond only CEFR level.
Sentence: `{sentence}`

C.2.3 Outputs from the Automatic Grammar Checker

The outputs from the automatic grammar checker are overly specific, identifying narrow error types such as “I told her (to) break a leg” or “this render (renders) the …”. To enable more effective analysis, we consolidated similar low-level errors into broader categories. For instance, the category “Omission of a Preposition” includes examples like “I told her (to) break a leg” and “It would be great (to) write a story.” The category “Mismatch between Article and Noun” captures cases such as “I like to use a pens and paper,” “I have received a 150 likes,” and “The cat is an animals.”

In total, we define 42 higher-level categories: “Confusion between effects and affects”, “Double negation”, “Gerund complement after psych/perception verb”, “Inappropriate formulaic closing”, “Incorrect existential agreement with plural noun”, “Incorrect passive voice usage”, “Incorrect pluralization after ‘either of’ ”, “Incorrect use of ‘if’ instead of ‘whether’ ”, “Incorrect use of gerund after ‘advise’ ”, “Incorrect verb usage with auxiliary”, “Mismatch between article and noun”, “Mismatch between noun and adjective”, “Mismatch between subject and verb”, “Missing complementizer ‘to’ after ‘allow” ’, “Missing determiner after quantifier”, “Misusage of irregular past tense verbs”, “Misuse of ‘have’ and ‘having’ ”, “Non-standard negation with ‘let’s’ ”, “Omission of a preposition”, “Omission of a verb”, “Omission of object pronoun”, “Omission of required articles”, “Omission of subject”, “Plural noun required after quantifier phrase”, “Redundant discourse marker usage”, “Redundant modal construction”, “Redundant phrase repetition”, “Redundant verb in question form”, “Singular form in fixed polite expression”, “Usage of ‘couple times’ instead of ‘a couple of times’ ”, “Usage of a plural noun when a singular form is required”, “Usage of a plural noun where a singular is required after ‘is there any’ ”, “Usage of a singular noun when a plural form is required”, “Usage of an adjective where an adverb is required”, “Usage of an auxiliary verb when unnecessary”, “Usage of an incorrect past participle form”, “Usage of first-person subject with ‘according to’ ”, “Usage of passive voice when active voice is required’ ”, “Usage of plural auxiliary ‘do’ with singular subject ‘anyone’ ”, “Use of ‘much’ with countable noun”, “Use of continuous aspect with stative verbs”, “Use of plural noun with each/every.”

C.2.4 L1-Specific Features

The following are the extracted features categorized by L1.

- Arabic: Usage of a plural noun where a singular is required after ‘is there any’, Incorrect passive voice usage, Usage of ‘couple times’ instead of ‘a couple of times’, Omission of a preposition, Mismatch between article and noun, Omission of a verb, Usage of a singular noun when a plural form is required, Omission of subject, Missing determiner after quantifier, Mismatch between article and noun
- Chinese-Mandarin: Usage of plural auxiliary ‘do’ with singular subject ‘anyone’, Inappropriate formulaic closing, Mismatch between subject and verb, Singular form in fixed polite expression, Omission of subject, Usage of an incorrect past participle form, Mismatch between article and noun, Incorrect existential agreement with plural noun, Usage of passive voice when active voice is required
- French: Non-standard negation with ‘let’s’, Usage of ‘couple times’ instead of ‘a couple of times’, Redundant verb in question form, Misuse of ‘have’ and ‘having’, Usage of a plural noun where a singular is required after ‘is there any’, Use of plural noun with each/every, Gerund complement after psych/perception verb, Omission of a preposition, Omission of a verb, Usage of first-person subject with ‘according to’
- German: Incorrect passive voice usage, Usage of ‘couple times’ instead of ‘a couple of times’, Misuse of ‘have’ and ‘having’, Gerund complement after psych/perception verb, Omission of a preposition, Incorrect verb usage with auxiliary, Misusage of irregular past tense verbs, Use of ‘much’ with countable noun, Usage of an adjective where an adverb is required, Incorrect use of gerund after ‘advise’
- Italian: Incorrect use of ‘if’ instead of ‘whether’, Usage of ‘couple times’ instead of ‘a couple of times’, Usage of a plural noun where a singular is required after ‘is there any’, Redundant discourse marker usage, Incorrect pluralization after ‘either of’, Gerund complement after psych/perception verb, Use of plural noun with each/every, Usage of a singular noun when a plural form is required, Omission of a verb, Misusage between ‘not’ and ‘never’
- Japanese: Use of continuous aspect with stative verbs, Mismatch between noun and adjective, Redundant modal construction, Usage of a singular noun when a plural form is required, Omission of a preposition, Gerund complement after psych/perception verb, Missing determiner after quantifier, Plural noun required after quantifier phrase, Omission of required articles, Omission of object pronoun
- Portuguese: Omission of a preposition, Omission of subject, Gerund complement after psych/perception verb, Usage of an auxiliary verb when unnecessary, Usage of a singular noun when a plural form is required, Missing complementizer ‘to’ after ‘allow’, Singular form in fixed polite expression, Redundant phrase repetition, Double negation, Incorrect existential agreement with plural noun
- Russian: Redundant verb in question form, Mismatch between article and noun, Misusage of preposition, Mismatch between subject and verb, Omission of a verb, Omission of subject, Missing complementizer ‘to’ after ‘allow’, Omission of a preposition, Redundant verb, Redundant preposition
- Spanish: Non-standard negation with ’let’s’, Incorrect pluralization after ‘either of’, Mismatch between article and noun, Omission of subject, Omission of a preposition, Incorrect verb usage

Table 8: Prompt for generating transformation guidelines

System:
You are a linguistic expert. I am a student trying to understand the given linguistic feature and transform a sentence reflecting the feature. As an expert, first explain the linguistic characteristics of the given linguistic feature. Then, outline detailed steps to transform a given sentence to reflect the characteristic, breaking the process into two main phases: Qualification and Application. Qualification refers to steps that identify if the linguistic feature is applicable to the given sentence in yes/no question format where answers to all questions should be 'yes' if feature dialect is applicable. Here, questions in qualification should be strictly related to lexical rules and should not ask for any decisions asking for significance or emotions. The questions should also not include questions about context, culture, or metaphors. Application refers to the action items that a model should take in order to reflect the given linguistic feature. All questions and action items should strictly be related to lexicon. All questions and action items should not include context, culture, or metaphor where answers might differ by people such as determining significant role or emotional attachment.

User:
Linguistic Feature: She/her used for inanimate referents.
Description: Pronouns, pronoun exchange, nominal gender.
Examples: The boat I had, was a seventy-two foot boat., She was built in Joneses-Slip in nineteen-fifty-five by my father., She's a nice bike., I left the boat there as she was., This is another pot and saucer., A bit dusty! You see that one isn't exactly glazed proper, burnt proper is she?, She's a twenty point five metre sloop again designed by Kel Steinman an Australian, So the Bransfield is the BAS ship then? Yeah she was the sort of kingpin until they brought the James Clark Ross

Assistant:
Linguistic Characteristic: She/Her Used for Inanimate Referents
The phenomenon of using 'she' or 'her' for inanimate objects is primarily influenced by cultural associations, personification, or grammatical gender in certain languages. In English, it is less common but can occur in specific contexts, such as referring to ships, cars, or countries, often reflecting affection, personification, or historical conventions. This use can evoke an emotional connection or imply a particular viewpoint about the inanimate referent.

Steps to Transform a Sentence: To transform a given sentence to reflect the use of 'she/her' for inanimate referents, we can break the process down into two main phases: Qualification and Application.
Qualification
1. Is there an inanimate referent in the sentence?
2. Does the original sentence contain a pronoun that can be replaced with 'she' or 'her'?
If the answers to all relevant questions are "Yes," then the linguistic feature is applicable.
Application
1. Identify the inanimate referent in the sentence that will be modified.
2. Replace any pronouns referring to the inanimate referent with "she" or "her."

User:
Linguistic Feature: `{linguistic feature}`
Description: `{description of linguistic feature}`
Examples: `{examples of linguistic feature}`

with auxiliary, Usage of a singular noun when a plural form is required, Missing Determiner after Quantifier, Redundant verb, Misusage of article in uncountable noun

- Turkish: Confusion between effects and affects, Usage of first-person subject with 'according to', Usage of a singular noun when a plural form is required, Omission of a preposition, Missing complementizer 'to' after 'allow', Omission of subject, Usage of a plural noun when a singular form is required, Missing determiner after quantifier, Mismatch between article and noun, Redundant adverb

Table 9: Examples of transformation guidelines.

Feature: Myself/meself instead of I in coordinate subjects **Qualification**: 1. Is there a coordinate subject in the sentence? A coordinate subject is formed when two subjects are joined by a conjunction like ‘and’ or ‘or’. 2. Does the coordinate subject include ‘I’? If the answers to all relevant questions are ‘Yes’, then the linguistic feature is applicable. **Application**: 1. Identify the coordinate subject in the sentence that includes ‘I’. 2. Replace ‘I’ with ‘myself’ in the coordinate subject.
Feature: Omission of Required Articles **Qualification:** 1. Does the sentence contain a noun that requires an article (‘a’, ‘an’, or ‘the’) for grammatical correctness or clarity? 2. Is the noun countable and in singular form, or does it refer to something specific that needs ‘the’? If the answers to all relevant questions are ‘Yes’, then the linguistic feature is applicable. **Application:** 1. Identify the noun(s) that require an article for grammatical correctness. 2. Remove the article (‘a’, ‘an’, or ‘the’) preceding the noun or leave the noun without any article.

C.3 Transformation Guideline Generation

We use `gpt-4-0613` to generate transformation guidelines via one-shot prompting, with a temperature of 0.8 and top-p sampling set to 0.95. The model is provided with the name of the linguistic feature, a brief description, and representative examples. It is then instructed to (1) describe the linguistic characteristics of the feature, and (2) outline a step-by-step transformation procedure consisting of two phases: *Qualification*, which checks whether the feature applies to a given sentence, and *Application*, which modifies the sentence accordingly.

We emphasize that the transformation process should focus strictly on lexical rules, avoiding subjective elements such as emotional or cultural interpretation, metaphor, or judgments of significance. The full prompt used for generating transformation guidelines is shown in Table 8, and examples of the resulting guidelines are presented in Table 9.

C.4 Transforming into English Varieties

C.4.1 Transformation of Vocabulary into Target CEFR Levels

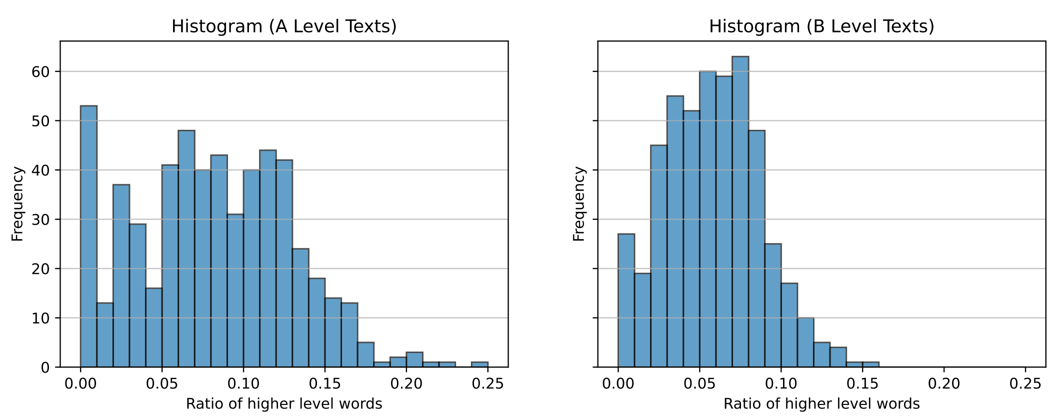

Figure 3: Histograms of distributions of higher-level word usage in CEFR A and CEFR B texts.

To ensure that the transformed outputs for ESL English varieties reflect realistic proficiency levels, we incorporated a vocabulary substitution step guided by CEFR-level annotations. To acknowledge that ESL learners often know a small fraction of advanced words even at lower proficiency levels, we first analyze a CEFR-labeled English text dataset to find out the ratio of higher-level words used by lower CEFR proficiency level learners, as mentioned in Section 3.3. Figure 3 presents the distribution of higher-level vocabulary in the dataset (*e.g.*, , B or C level words in A level texts). Notably, for both target levels A and B, at least 90% of the samples contain no more than 15% of vocabulary from higher CEFR levels than the designated target level. This empirical finding motivated our decision to allow up to 15% of higher-level vocabulary in transformed outputs. This threshold balances fidelity to learner-level constraints with linguistic realism, acknowledging that ESL learners often know a small fraction of advanced words even at lower proficiency levels.

Table 10: Vocabulary pseudo-label prompt.

System: You are an expert in classifying vocabulary into CEFR levels. Given a single word, classify it into its appropriate CEFR level when used with its most common definition. If it is a proper noun, answer with A1. Answer only with one of the following: A1, A2, B1, B2, C1, C2.

User: `{word}`

Table 11: Vocabulary transformation prompt.

System: You are an expert in transforming vocabulary of higher CEFR levels to level `{target_level}`. You are given higher level words that appear in the question: `{words_to_transform}`. Please replace at least `{min_transform_words}` words with synonyms in level `{target_level}`.

User: `{question_text}`

Table 10 shows the prompt used for finding pseudo-labels for words without a CEFR label in the Oxford vocabulary lists, and Table 11 presents the prompt used for transforming higher-level vocabulary in a sentence to a target level. The value of `min_transform_words` is set to 15% of the total word count in `question_text` and serves as the threshold for permitted higher-level words. Table 12 presents the transformation success rates by CEFR level and dataset, showing how often our pipeline was able to produce outputs that met CEFR-level vocabulary constraints while preserving semantic equivalence.

Table 12: Number and ratio of valid vocabulary transformations by dataset.

Dataset	Size	Target CEFR	Valid Transf.	Transf. Ratio
MMLU	14042	A	7246	51.6%
		B	11970	85.2%
GSM8K	1319	A	1219	92.4%
		B	1315	99.7%
ARC	1172	A	774	66.0%
		B	1132	96.6%
HellaSwag	10042	A	7593	75.6%
		B	9903	98.6%
TruthfulQA	817	A	623	76.3%
		B	781	95.6%
WinoGrande	1267	A	945	74.6%
		B	1247	98.4%

C.5 Prompts used for Transformation

Table 13 presents the one-shot prompt used to transform a Standard American English (SAE) sentence s into a target variety using the feature transformation model T. Each transformation is guided by

Table 13: Prompt for transforming into varieties.

System: Your task is to rephrase the given sentence by following the guideline.
{transformation guideline}

1. **Qualification**:
- Answer the qualification questions for the linguistic feature with either "yes" or "no."
- Answer the questions in a very strict manner.
- Proceed to the next step only if **all** answers are "yes."
- Otherwise, stop in qualification phase with generating '**Transformed Sentence:** (No change)'.

2. **Application**:
- Make only the **necessary changes** to apply the linguistic feature, ensuring no loss of information.
- Provide the final transformed sentence, adhering strictly to the format and structure of the given example.

Mandatory
- Proceed to Application only if all answers to the qualification questions are 'yes'.
- Preserve the structure of the original sentence as much as possible with no information loss.
- Follow the guideline, not considering standard English grammar.
- Final sentence should start with '**Transformed Sentence:**' either with sentence of (No change).

User: **Original Sentence**: `{example sentence}`

Assistant: `{example output}`

User: **Original Sentence**: `{SAE written sentence}`

Table 14: Prompt for semantic check.

User: Determine whether the meaning of Sentence 1 is significantly altered or lost in Sentence 2.

Consideration
- All keywords from Sentence 1 should be in Sentence 2.
- All numbers in Sentence 1 should match with Sentence 2.
- Focus on core information only.
- Ignore grammar; it is not a factor for consideration.
- Missing or incorrect prepositions should not be considered.
- Ignore repetition of phrases. Repetition is not a factor for consideration.
- Base your decision solely on whether essential information is missing.

Respond with either 'yes' or 'no' only.

Sentence 1: `{SAE written sentence}`
Sentence 2: `{transformed sentence}`

Answer:

a feature-specific guideline and example. The model is instructed to follow the guideline strictly, preserving the structure and core meaning of the original sentence while disregarding grammatical correctness.

To ensure semantic fidelity, we employ a semantic checker model S using a zero-shot prompt, as shown in Table 14. The verification process emphasizes the preservation of key content elements such as keywords, numerical information, and core propositions, while ignoring minor grammatical deviations, including incorrect or missing prepositions and redundancy.

Table 15: Average number of features applied per sample and proportion of transformed samples in dialect.

	MMLU	ARC	TruthfulQA	GSM8K	Hellaswag	WinoGrande
AAVE	1.12 / 61.8%	1.17 / 65.1%	0.76 / 45.0%	0.80 / 54.7%	2.01 / 87.4%	2.06 / 88.4%
AppE	1.53 / 70.6%	1.14 / 64.9%	1.08 / 60.5%	1.11 / 63.4%	2.26 / 88.9%	2.70 / 96.5%
AuE	0.80 / 65.3%	0.76 / 64.8%	0.49 / 41.5%	0.40 / 33.0%	0.91 / 66.5%	1.60 / 96.8%
AusVE	0.95 / 57.5%	0.76 / 50.7%	0.78 / 57.8%	1.05 / 70.5%	1.53 / 82.9%	1.63 / 91.9%
BahE	2.63 / 70.5%	1.94 / 53.7%	1.76 / 63.4%	2.91 / 76.6%	3.20 / 83.9%	6.22 / 99.5%
EAngE	3.54 / 87.7%	3.08 / 86.1%	2.87 / 90.0%	3.75 / 90.2%	4.58 / 95.9%	5.94 / 99.8%
IrE	2.67 / 91.0%	2.92 / 95.0%	2.49 / 87.8%	1.80 / 78.8%	4.82 / 98.9%	4.53 / 100.0%
Manx	1.86 / 86.8%	1.64 / 86.9%	1.57 / 80.7%	0.84 / 60.5%	2.57 / 95.8%	3.22 / 98.3%
NE-Eng	0.70 / 59.6%	0.77 / 70.5%	0.43 / 38.8%	0.58 / 54.7%	1.43 / 89.9%	1.05 / 77.0%
NZE	2.07 / 84.7%	2.12 / 88.2%	1.48 / 70.3%	2.15 / 85.8%	3.10 / 97.3%	3.48 / 99.4%
NfE	4.17 / 95.4%	3.98 / 96.4%	3.31 / 92.5%	4.3 / 96.7%	5.55 / 98.9%	7.63 / 99.9%
OzE	2.50 / 86.6%	2.73 / 91.9%	2.17 / 85.8%	2.75 / 89.8%	3.59 / 96.6%	4.07 / 99.2%
SE-AmE	2.50 / 79.9%	2.19 / 70.8%	2.03 / 79.1%	2.98 / 84.9%	3.65 / 91.4%	4.72 / 99.6%
SE-Eng	0.20 / 17.4%	0.14 / 13.2%	0.07 / 6.6%	0.22 / 19.7%	0.26 / 22.9%	0.30 / 25.7%
SW-Eng	0.90 / 66.3%	0.77 / 62.9%	0.55 / 43.6%	0.33 / 30.0%	0.84 / 64.4%	1.67 / 96.9%
ScE	1.15 / 69.8%	1.06 / 67.9%	1.06 / 63.5%	0.76 / 51.2%	1.20 / 70.3%	2.05 / 97.8%
TdCE	0.94 / 44.9%	0.69 / 35.9%	0.47 / 31.1%	1.12 / 54.7%	2.17 / 85.8%	2.10 / 92.9%
WeE	2.27 / 90.1%	2.11 / 89.9%	2.51 / 97.1%	1.08 / 61.0%	1.81 / 83.5%	3.01 / 98.7%

Table 16: Average number of features applied per sample and proportion of transformed samples in ESL English.

		MMLU	ARC	TruthfulQA	GSM8K	Hellaswag	WinoGrande
A	ar	2.65 / 96.6%	2.77 / 98.8%	2.05 / 92.6%	3.06 / 99.5%	2.88 / 98.2%	2.87 / 99.8%
	de	2.17 / 93.4%	2.30 / 94.8%	1.92 / 91.8%	2.15 / 94.9%	2.88 / 96.0%	2.98 / 99.7%
	es	3.15 / 97.1%	3.50 / 99.7%	2.74 / 97.1%	3.53 / 99.3%	3.55 / 98.3%	3.63 / 99.6%
	fr	1.00 / 84.6%	0.99 / 86.6%	0.83 / 74.8%	1.15 / 87.2%	1.16 / 87.6%	1.11 / 92.8%
	it	1.03 / 80.8%	1.05 / 87.5%	0.75 / 68.1%	1.20 / 87.3%	1.33 / 89.6%	1.19 / 87.7%
	ja	3.21 / 96.5%	3.41 / 98.8%	2.54 / 94.2%	3.20 / 98.1%	3.93 / 98.1%	3.83 / 100.0%
	pt	2.92 / 98.1%	3.07 / 99.5%	2.89 / 99.4%	3.30 / 99.8%	3.27 / 98.3%	3.36 / 99.9%
	ru	3.02 / 97.5%	3.28 / 99.7%	2.85 / 99.0%	3.53 / 99.5%	3.33 / 98.6%	3.56 / 99.9%
	tr	2.94 / 96.9%	3.08 / 97.9%	2.34 / 92.9%	3.29 / 98.0%	3.18 / 97.6%	3.22 / 99.9%
	zh	1.63 / 88.0%	1.67 / 90.6%	1.23 / 83.6%	2.02 / 93.6%	1.77 / 89.6%	1.84 / 93.2%
B	ar	2.83 / 96.4%	2.82 / 98.2%	2.09 / 91.5%	3.15 / 98.5%	2.84 / 98.8%	2.89 / 99.2%
	de	2.09 / 92.3%	2.01 / 91.9%	1.95 / 91.8%	1.96 / 91.5%	2.54 / 94.7%	2.98 / 99.7%
	es	3.27 / 97.4%	3.43 / 98.9%	2.89 / 97.4%	3.59 / 98.9%	3.30 / 98.8%	3.51 / 99.9%
	fr	0.97 / 83.0%	0.91 / 82.3%	0.79 / 70.2%	1.05 / 84.2%	1.01 / 77.8%	1.11 / 89.2%
	it	0.93 / 73.4%	0.87 / 73.6%	0.65 / 56.2%	1.07 / 79.4%	1.30 / 87.8%	1.21 / 88.5%
	ja	3.31 / 96.6%	3.16 / 98.0%	2.53 / 92.4%	3.06 / 97.9%	3.52 / 98.3%	3.82 / 99.8%
	pt	2.95 / 98.2%	3.05 / 99.0%	2.91 / 99.0%	3.25 / 99.5%	2.94 / 97.2%	3.29 / 99.8%
	ru	3.15 / 97.5%	3.33 / 99.6%	3.01 / 98.6%	3.59 / 99.3%	3.02 / 97.7%	3.52 / 99.9%
	tr	3.06 / 96.2%	2.96 / 96.8%	2.24 / 88.2%	3.26 / 98.5%	3.00 / 98.3%	3.19 / 99.9%
	zh	1.83 / 93.3%	1.94 / 97.3%	1.44 / 88.3%	2.2 / 97.8%	1.97 / 96.1%	2.04 / 96.6%

C.6 Transformation Ratio

Tables 15 and 16 report the average number of features applied per sample and the overall proportion of transformed samples for dialect and ESL English, respectively, as discussed in Section 3.4. Consistent with the results presented in the main paper, ESL English exhibits a higher transformation rate and a greater average number of features applied per sample compared to dialects.

C.7 Human Evaluation

Human annotators were shown one sample at a time, with a total of 150 samples randomly shuffled, 25 from each model. For each sample, annotators answered two binary (yes/no) questions: (Q1) whether the model correctly followed the Qualification and Application steps specified in the transformation

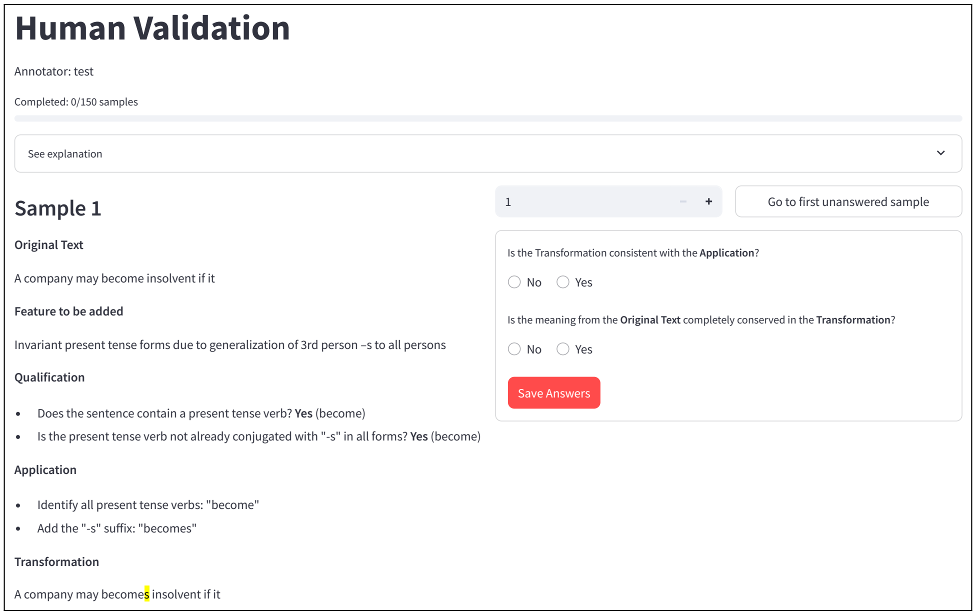

Figure 4: Interface used for human evaluation.

guideline, and (Q2) whether the transformed sentence preserved the original meaning. The interface presented to annotators is shown in Figure 4. A sample was considered valid if it received majority approval from the annotators.

D Experiments

D.1 Experiment Setting

We evaluated the transformed datasets on seven state-of-the-art models: Qwen2.5-72B-Instruct [63], DeepSeek-R1-Distill-Llama-70B [12], LLaMA-3.3-70B-Instruct [23], Gemini 2.0 Flash [20], Gemini 2.5 Pro [19], GPT-4o-mini [47], and o4-mini [49].[13] We set the maximum number of generated tokens to 2048 and conducted all experiments in a zero-shot setting. The system prompt used was: "Do not reason for too long. If the question is a multiple choice question, answer with the option letter. If none of the given options match, you may guess or say 'none of the above.' Start your final sentence with 'The answer is '." To extract the model's prediction, we parsed the output beginning from the phrase "The answer is", using the subsequent text as the final answer.

D.2 Full Experiment Analysis

Figures 5 and 6 present the full analysis results across all datasets, corresponding to the analysis in Section 4.2. Consistent with the findings in the main paper, we observe a positive correlation between linguistic distance from Standard American English (SAE) and performance degradation, although the strength of this relationship varies across datasets. In ESL English, despite some deviations, performance drop generally increases with the difficulty level of the English variety.

[13]Model versions: `gemini-2.5-pro-exp-03-25`, `gpt-4o-mini-2024-07-18`, `o4-mini-2025-04-16`

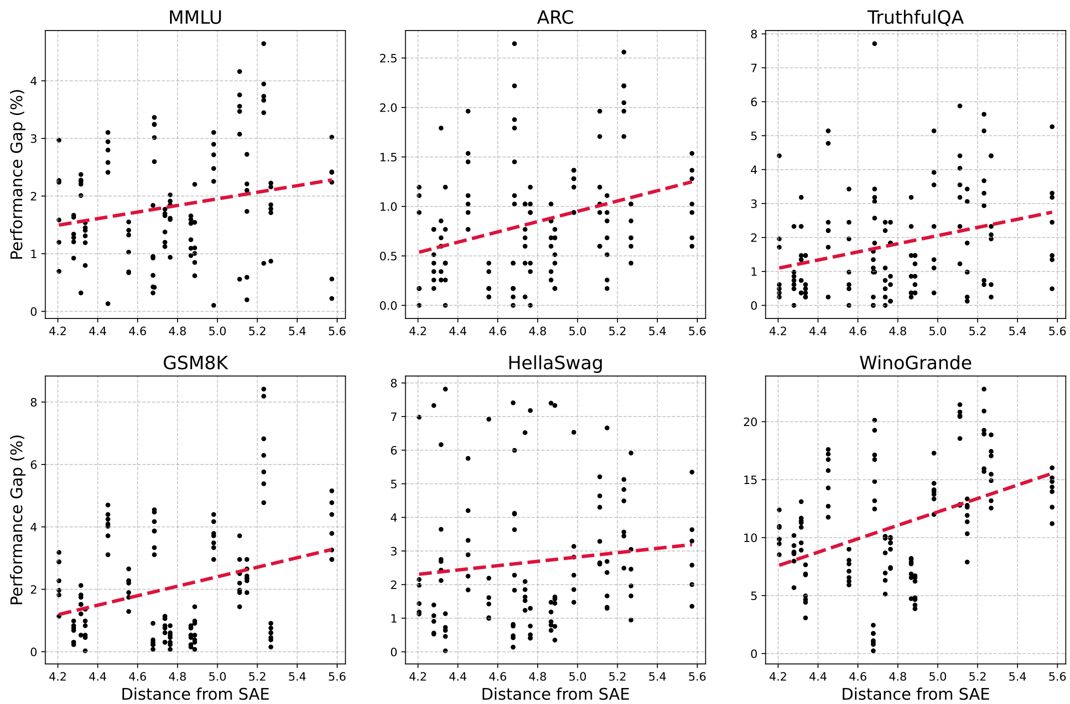

Figure 5: Correlation between linguistic distance and model performance degradation.

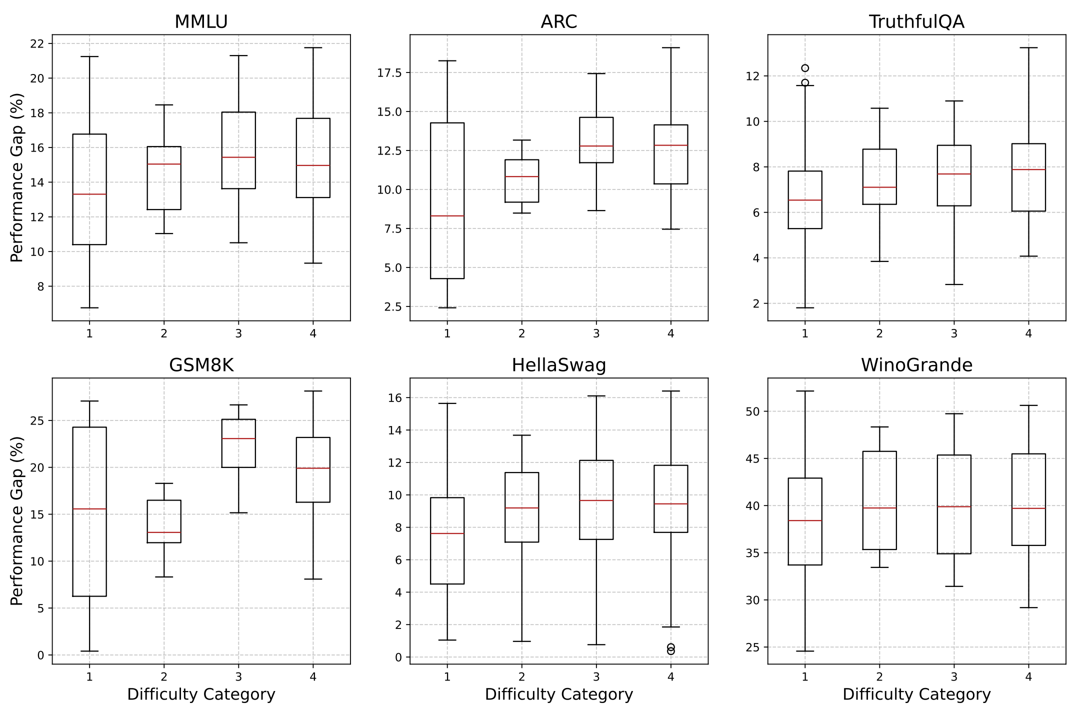

Figure 6: Boxplot by difficulty category and model performance degradation.