Feature Vulnerability in Automated Essay Scoring

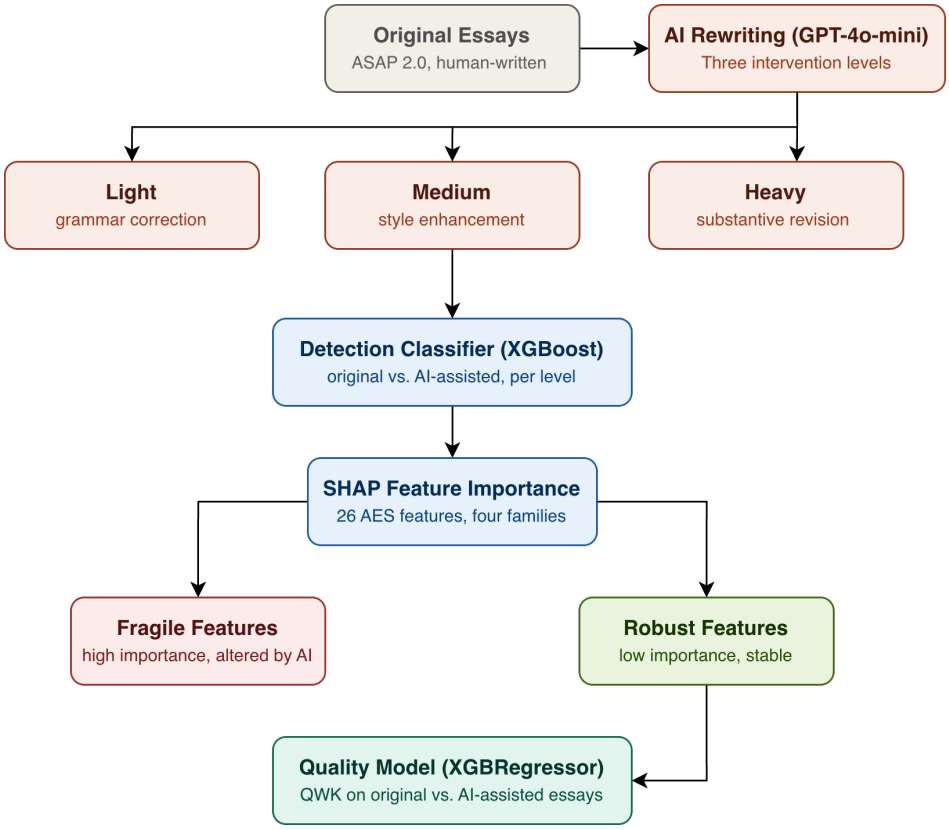
A Detection-Based Analysis of AI Writing Assistance

submitted in partial fulfillment for the degree of master of science

Zakaria Hader
16347005

Master Information Studies
data science
Faculty of Science
University of Amsterdam

Submitted on 26.06.2026



	UvA Supervisor
Title, Name	Dr. Jelke Bloem
Affiliation	UvA Supervisor
Email	j.bloem@uva.nl

Abstract

Automated essay scoring (AES) relies on linguistic features developed on human-written text. Widespread AI writing assistance raises a validity concern: AI may alter the features AES depends on, so scores increasingly reflect AI access rather than writing ability. AES and AI-detection research have developed independently, leaving open whether AES features remain reliable under AI assistance. This thesis proposes a *detection as diagnosis* framework: a classifier distinguishing AI-assisted from human-written essays identifies which AES features are *fragile* (high detection importance) versus *robust* (low importance, stable). Using ASAP 2.0, essays were rewritten by GPT-4o-mini at three intensities, and classifiers detected each. Detection accuracy increased monotonically with intensity (AUC 0.719 to 0.999). SHAP analysis identified seven fragile features, dominated by average word length, and seven robust features, stable across folds. A quality model using only robust features did not consistently exceed the conventional QWK ≥ 0.70 threshold, but degraded less under heavy intervention than a model using all features. Feature fragility under AI assistance can be diagnosed via detection classifiers, and restricting to non-discriminative features improves stability, though not absolute reliability, of quality assessment under AI-assisted writing.

Keywords

Automated Essay Scoring, AI Writing Assistance, Feature Robustness, SHAP Analysis, AI-Generated Text Detection

Github Repository

https://github.com/Zakaria0531/master-thesis

1 Introduction

The assessment of academic writing quality has increasingly relied on automated metrics and natural language processing techniques. Automatic essay scoring (AES) systems and text complexity measures enable scalable, consistent evaluation of student writing and have been validated against human ratings in numerous educational contexts [10]. These systems typically extract linguistic features including lexical sophistication, syntactic complexity, and discourse coherence, with tools like Coh-Metrix providing computational analysis of cohesion across multiple linguistic levels [7].

However, the recent widespread adoption of AI-based writing assistants introduces a fundamental validity challenge to these assessment tools. Grammar correction tools like Grammarly, style enhancement features in word processors, and particularly large language models (LLMs) such as ChatGPT are now routinely used by students in academic settings. These AI assistants systematically improve the superficial linguistic features that many automated metrics rely upon, creating a critical measurement problem: existing metrics may no longer accurately assess a student's genuine writing ability, but rather their access to and proficiency with AI assistance tools.

Simultaneously, the detection of AI-generated and AI-assisted text has emerged as an active research area. Detection approaches analyze statistical patterns, train discriminative classifiers, and examine linguistic markers that distinguish human from machine text [4]. However, this detection literature and the AES literature have remained largely disconnected despite addressing complementary aspects of the same problem: understanding which features of text are most affected by AI involvement.

Bridging these domains is scientifically relevant in three key aspects. First, identifying which AES features are most discriminative for detecting AI assistance reveals which features are most vulnerable to manipulation, providing empirical evidence regarding metric robustness. Second, features that prove non-discriminative for detecting AI assistance are candidates for robust quality assessment, as they appear less sensitive to AI-induced changes in the text. Third, this integration enables development of assessment frameworks that maintain validity in educational contexts where AI assistance is increasingly prevalent.

The research gap lies at the intersection of these two literatures. While AES research assumes human authorship and focuses on predicting quality scores, detection research assumes binary categories and focuses on identifying AI involvement. There has been little systematic investigation into whether features used in automated essay scoring remain reliable when writing is assisted by AI tools. Furthermore, there is limited understanding of how different types of AI interventions (grammar correction versus substantive revision) affect different categories of linguistic features.

This thesis addresses the following research question:

RQ: To what extent can automated essay scoring features distinguish AI-assisted from human-written text, and which features remain robust for quality assessment under AI assistance?

This main research question decomposes into three sub-research questions that progressively build from detection to robust assessment:

SRQ1: How well can binary classifiers distinguish original student essays from AI-assisted variants across three intervention levels (grammar correction, style enhancement, substantive revision)?

This question examines whether AI assistance produces measurable changes in the feature space used by automated essay scoring systems. If AI-assisted essays can be reliably distinguished from original essays, this indicates that AI intervention systematically alters the linguistic characteristics on which AES models rely. I expect detection performance to increase with intervention intensity, as more extensive revisions are likely to introduce larger feature shifts.

SRQ2: Which AES features show the highest importance for detecting AI assistance versus the lowest importance?

This question identifies which AES features are most affected by AI assistance and which remain comparatively stable. Features that strongly contribute to detection can be considered fragile, whereas features with little discriminative value may be more robust to AI-induced changes. I expect surface-level and readability-related features to be more fragile than discourse and syntactic features.

SRQ3: Can quality assessment models using only robust features show less performance degradation than all-feature

models when applied to AI-assisted essays? This question evaluates whether restricting assessment to robust features improves the stability of quality prediction under AI-assisted writing. If models based on robust features experience less performance degradation than models using the full feature set, this would suggest that more reliable assessment remains possible even when students use AI writing assistance. Assessment quality is considered both in terms of stability under AI intervention and in relation to the conventional AES acceptability threshold of QWK ≥ 0.70 [14], providing insight into whether robust features can support practically useful assessment.

This research contributes to the literature in several ways. Methodologically, it bridges the automated essay scoring and AI detection literatures through a detection as diagnosis framework that uses classifier feature importance to identify vulnerabilities in commonly used AES metrics. Empirically, it provides a systematic characterization of how different forms of AI assistance affect distinct categories of linguistic features. This analysis is enabled by a dataset in which each original essay is transformed into multiple AI-assisted variants representing grammar correction, stylistic enhancement, and substantive revision.

2 Related Work

The assessment of academic writing has traditionally relied on automated metrics that assume human authorship, while recent advances in AI writing assistance have created unprecedented challenges for these evaluation systems. Although substantial research exists in automated essay scoring and emerging work addresses AI-generated text detection, the intersection of these domains remains largely unexplored. Specifically, there is limited systematic investigation into which AES features are most useful for detecting AI assistance, and conversely, which features remain robust for quality assessment when AI tools are used. This section reviews the relevant literature across three key areas: automated essay scoring systems, AI-assisted text detection, and the emerging challenges of assessment validity in AI-assisted contexts.

2.1 Automated Essay Scoring Systems and Features

Automated essay scoring has evolved from simple surface features to sophisticated neural approaches. Ke and Ng [5] provides a comprehensive overview of AES systems, categorizing approaches into feature-based methods (extracting linguistic features and training ML models) and neural methods (using pre-trained language models). Traditional feature based systems rely on a broad range of linguistic features, including lexical diversity, syntactic complexity, grammatical error counts, readability indices, and discourse cohesion metrics. McNamara et al. [7] developed one of the most comprehensive feature extraction tools, providing over 100 indices across multiple linguistic levels. More recent approaches employ neural architectures such as recurrent neural networks and transformer based models to directly predict essay scores from text representations [12]. While these models achieve strong predictive performance, they may still rely on surface level cues and can be vulnerable to adversarial manipulation.

Critically, these AES systems were developed and validated on human-written text, with no systematic evaluation of their behaviour when applied to AI-assisted writing. Shermis [10] established performance benchmarks on the ASAP dataset, which will be used as the primary corpus. The feature categories commonly used in AES systems (surface-level, readability, discourse, syntactic) form the basis for my investigation into which features remain robust under AI assistance. The Automated Student Assessment Prize (ASAP) competition further demonstrated that machine learning approaches can achieve scoring performance comparable to human raters.

This thesis extends this literature by systematically evaluating how established AES features behave when essays are modified through AI assistance, using detection to identify feature vulnerabilities. As such, this body of work primarily motivates SRQ2 and SRQ3, which investigate feature robustness and its implications for quality assessment under AI-assisted writing.

2.2 AI-Generated and AI-Assisted Text Detection

The detection of AI-generated text has become an active research area following the release of powerful LLMs. Gehrmann et al. [4] developed statistical detection methods based on analysing token probability distributions. More recent work [8] demonstrates that zero-shot detection is possible using perplexity and related metrics that measure how predictable text is under a language model. However, detection performance generally degrades as language models become more capable. OpenAI [9] acknowledges that their own classifier achieves only 26% true positive rate at 9% false positive rate, highlighting detection difficulty. The tool was subsequently discontinued in July 2023 due to low accuracy.

Recent benchmark efforts [3] provide large-scale evaluation frameworks for machine-generated text detection, showing that detection remains challenging across domains and models. Liang et al. [6] found that non-native English speakers are disproportionately flagged as AI-generated, raising concerns about detection fairness.

However, this detection literature focuses primarily on distinguishing fully AI-generated text from human-written text, treating generation as a binary category. Limited work exists on detecting AI assistance (where humans use AI to enhance their own writing) or on connecting detection features to quality assessment metrics. Weber-Wulff et al. [13] found that detection accuracy drops significantly when AI-generated text is lightly edited, suggesting that assistance is harder to detect than full generation.

Unlike prior detection studies, this thesis focuses on AI-assisted rather than fully AI-generated text and uses detection as a diagnostic mechanism for identifying vulnerable AES features. As such, this literature primarily motivates SRQ1 and SRQ2.

2.3 Challenges of Assessment in AI-Assisted Writing Contexts

Recent studies document the widespread use of AI writing tools in education and raise concerns about assessment validity. Sullivan et al. [11] report that 30-40% of students acknowledge using ChatGPT for academic work. Yan [15] shows that AI assistance

measurably improves surface-level writing features (grammar, vocabulary) but with unclear effects on deeper learning.

The approach taken here addresses this gap by providing empirical evidence about feature-level robustness and developing assessment approaches that maintain validity despite AI assistance, rather than attempting to prevent or detect all AI use. As such, this literature primarily motivates SRQ3, which investigates whether robust features can support more stable quality assessment under AI-assisted writing.

3 Methodology

This study employs a detection-based approach to determine which automated essay scoring (AES) features are fragile (easily shifted by AI writing assistance) and which are robust (remaining stable regardless of intervention). The core insight is that a classifier trained to distinguish AI-assisted from human-written essays implicitly learns which features change: high-importance features are fragile, low-importance features are robust. This approach is novel relative to prior AES work, which typically treats feature sets as fixed. Here, the detection task serves as a diagnostic lens on feature stability. While feature importance does not directly measure causal sensitivity to AI intervention, features that consistently contribute to distinguishing original from AI-assisted essays provide a practical proxy for identifying which metrics are most affected by AI-induced changes. The subsections below describe the dataset, intervention protocol, feature extraction strategy, classification design, and validity considerations, with each component mapped to a specific sub-research question in Section 3.6.

3.1 Dataset

I use the Automated Student Assessment Prize 2.0 (ASAP 2.0) dataset [2], which contains 24,728 student essays across seven prompts, each scored by a single human rater on a scale of 1–6 (with one prompt capped at 5). Essays have a mean length of 363 words (SD = 148, median = 338), with prompt-level medians ranging from 263 words (*A Cowboy Who Rode the Waves*) to 430 words (*Car-free cities*), confirming substantial prompt-dependent variation in length. Score distributions are concentrated in the mid-range (mean = 2.94, SD = 1.04, median = 3), with low and high scores occurring less frequently. Critically, word count alone explains approximately 50% of score variance ($R^2 = 0.50$), indicating that surface-level features are strongly predictive of quality ratings and may therefore be sensitive to AI-induced modifications. Figure 1 illustrates this relationship, showing that median word count increases monotonically with essay score, though substantial within-score variance indicates that length alone is not determinative. The dataset predates widespread LLM use, ensuring all source texts are genuine human writing with no AI contamination.

3.2 AI Intervention Protocol

To simulate real-world AI assistance across a range of intensities, I create three intervention levels using GPT-4o-mini. This model is selected over larger variants (e.g., GPT-4o, GPT-5) on cost grounds, and over open-source alternatives because API-based access with fixed decoding parameters (temperature=0, seed=42) provides the strongest reproducibility guarantee available for this model, independent of local infrastructure. Note that this does not guarantee fully deterministic outputs across API calls, as minor variation can still occur due to backend non-determinism. Each original essay is processed at all three levels, producing three AI-assisted variants per essay.

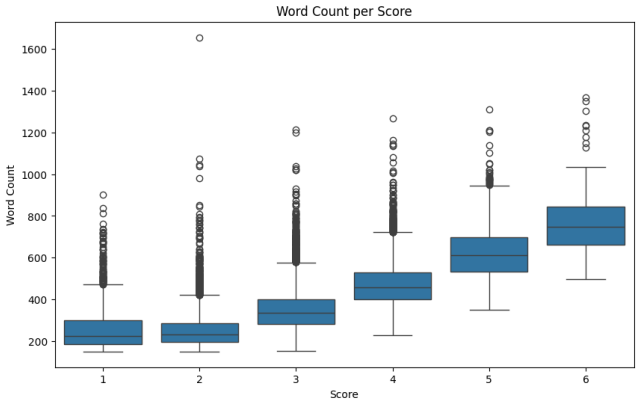


Figure 1: Word count by essay score on original essays. Median word count increases monotonically with score, but substantial overlap between adjacent score categories indicates that length alone does not determine quality ratings.

Level 1: Grammar Correction. Essays are processed with the system prompt:

"Fix only surface-level errors: spelling mistakes, grammar errors, and punctuation issues in the student essay below. Do not change word choice, sentence structure, or any ideas. Return only the corrected essay text with no commentary."

This simulates surface-level proofreading tools such as Grammarly.

Level 2: Style Enhancement. Essays are processed with the system prompt:

"Improve the student essay below by:
(1) correcting all spelling, grammar, and punctuation errors,
(2) improving sentence clarity and variety where needed, and
(3) replacing weak or imprecise word choices with stronger alternatives.
Do not add new arguments, examples, or ideas, and do not reorganize paragraphs. Return only the improved essay text with no commentary."

This represents intermediate writing assistance affecting lexical and syntactic surface form while preserving semantic content.

Level 3: Substantive Revision. The most aggressive intervention uses the system prompt:

"Substantially improve the student essay below while preserving its original arguments, and examples. Make comprehensive improvements to:
(1) grammar, spelling, and punctuation,
(2) vocabulary and academic register,
(3) sentence variety and fluency,
(4) paragraph structure and internal coherence, and
(5) logical flow and transitions between ideas.

The essay should sound like an advanced, polished version of the original student's essay, and not a replacement. Return only the revised essay text with no commentary."
This simulates comprehensive AI-assisted rewriting affecting discourse structure and argumentation.

The prompts were designed to form an intensity gradient across the three levels, grammar correction, style enhancement, and substantive revision, through explicit constraints on scope (e.g. "do not change word choice" at Level 1, "do not add new arguments, examples, or ideas, and do not reorganize paragraphs" at Level 2). Each prompt instructs the model to return only the revised essay text, with no additional commentary, to facilitate downstream feature extraction.

3.3 Feature Extraction

Features are extracted across four families, chosen because prior AES work [5, 7, 10] has shown these to capture complementary aspects of writing quality. The key hypothesis motivating feature selection is that surface-level and readability features are the most fragile, because AI assistants directly target vocabulary and syntactic complexity, while coherence and syntactic features depend on discourse-level organization that is harder to alter systematically and are therefore expected to be more robust.

Surface-level features capture lexical diversity, part-of-speech distributions, and basic length statistics. These are hypothesized to be highly fragile: even grammar correction (Level 1) can alter sentence length and lexical diversity by fixing run-ons or replacing repeated words.

Readability indices aggregate surface form into scalar grade-level or difficulty scores. These are expected to shift substantially at Level 2, where vocabulary sophistication is explicitly targeted.

Coherence features capture discourse-level organization through connective usage and lexical overlap between adjacent sentences. These are hypothesized to be moderately fragile at Level 3 (which reorganizes paragraphs) but robust at Level 1.

Syntactic features capture sentence-level structural complexity, including dependency depth, clause subordination, and voice. These depend on grammatical construction choices that are not directly targeted by surface-level edits, and are therefore hypothesized to be comparatively robust.

All features are computed using established libraries (spaCy for linguistic annotation, textstat for readability) and z-score standardized before classification. Table 4 lists the full set of 26 features by family.

3.4 Classification and Feature Importance Analysis

For each intervention level, I train a binary classifier to distinguish original essays from AI-assisted variants. I use gradient-boosted trees (XGBoost) as the primary model, with Random Forest as a secondary comparison. Tree-based models are preferred over neural or linear classifiers here because (a) they produce native feature importances that align directly with SHAP-based attribution and (b) they handle the dataset size (~25,000 essays, yielding ~75,000 AI-assisted variants) without requiring fine-tuning of pre-trained representations. This keeps the analysis interpretable and computationally feasible.

Data splitting. For each intervention level $k \in \{1, 2, 3\}$, a separate binary classifier is trained to distinguish original essays from level-k AI-assisted essays, on a 1:1 paired dataset (each original essay is paired with exactly one assisted variant at level k). The split is performed at the `essay_id` level using 5-fold `GroupKFold` cross-validation, ensuring each essay's original and assisted variant are assigned to the same fold and preventing identity leakage across splits. No separate held-out test set is used: the held-out fold in each CV iteration serves as the test set, and reported metrics are averaged across all five folds.

Evaluation. Detection performance is measured by AUC-ROC and macro-F1, both averaged across folds. Macro-F1 is used (rather than micro-F1) to weight each class equally, since `GroupKFold` can produce slight per-fold imbalance even with a 1:1 paired dataset. Feature importance is computed using SHAP global mean absolute values, averaged across cross-validation folds and intervention levels. The resulting fragility categorization is described in Section 3.6.

3.5 Reliability and Validity

Construct validity. The approach equates feature fragility with classifier-assigned importance. This is valid if the classifier's decision boundary reflects genuine feature shifts caused by the intervention, not confounds such as prompt-length effects or model-specific artifacts. The intervention protocol (Section 3.2) and the essay-level data split are the primary controls for this.

Internal validity. Essay-level cross-validation ensures the classifier never sees a variant of an essay it was trained on, preventing identity leakage. Manual inspection of intervention outputs across a range of essay types confirms that the prompts behave consistently, reducing the risk that observed feature shifts reflect prompt-specific artifacts rather than genuine intervention-level effects.

External validity. The ASAP 2.0 dataset's diversity (7 prompts, multiple genres) supports generalization to secondary-level academic writing in English. Results are not expected to generalize directly to university-level writing or non-English essays, which differ in genre conventions and baseline linguistic complexity.

Reproducibility. All code, feature extraction pipelines, and model configurations will be made publicly available. The ASAP 2.0 dataset is publicly accessible. Random seeds are fixed (42) for all stochastic processes, and GPT-4o-mini generation uses temperature 0 with a fixed seed to minimize stochastic variation in the intervention outputs.

3.6 Mapping Evaluation to Research Questions

SRQ1 (Detection accuracy across intervention types) is answered by training three separate binary classifiers, one per intervention level, each distinguishing original essays from level-k AI-assisted essays using XGBoost (with Random Forest as a secondary comparison). AUC-ROC and macro-F1 for each classifier are compared across the three intervention levels. The prediction is that Level 3 (substantive revision) is most detectable and Level 1 (grammar correction) least detectable. Deviations from this ordering would indicate that the intervention levels do not form a clean intensity gradient.

SRQ2 (Feature fragility analysis) is answered by computing, for each feature, a global SHAP importance score (`mean_shap_overall`): the mean absolute SHAP value is first averaged across validation samples within each cross-validation fold, then across the 5 folds to obtain a per-level mean, and finally across the three intervention levels (unweighted) to obtain a single global value per feature. Features are categorized by comparing this value against the first and third quartiles (Q_1, Q_3) of its distribution across all 26 features: features at or below Q_1 are categorized as robust, features at or above Q_3 as fragile, and the remainder as neutral. This data-driven thresholding yielded 7 robust and 7 fragile features.

SRQ3 (Quality assessment with robust features) is addressed by training an XGBoost regression model (XGBRegressor) to predict ASAP human scores, evaluated using Quadratic Weighted Kappa (QWK), the standard evaluation metric in AES research [5, 10]. Three feature subsets are compared: (a) all 26 extracted features, (b) only the features categorized as robust in SRQ2, and (c) only the features categorized as fragile in SRQ2. For each subset, a model is trained on Level-0 (original) essays only via 5-fold `GroupKFold` cross-validation grouped by `essay_id`, preventing identity leakage between original essays and their AI-assisted variants. The held-out Level-0 fold in each split serves as the test set, so no separate held-out set is used. Each fold's trained model is additionally applied, without retraining, to the AI-assisted essays at Levels 1–3, allowing QWK to be computed separately at each intervention level (L0–L3) and degradation in predictive performance to be tracked as intervention intensity increases. Performance is assessed against the operational QWK threshold of 0.70 [14], the standard benchmark for automated scoring deployment. The hypothesis is that the robust-feature model degrades less across intervention levels than the all-features or fragile-features models, demonstrating that non-fragile features provide a more stable basis for quality assessment under AI assistance, even if absolute performance does not reach the deployment threshold.

Full hyperparameter settings for all classification and regression models are reported in Appendix C (Table 5).

4 Results

This section reports the empirical results for the three sub-research questions in turn: detection accuracy across intervention levels (SRQ1), the resulting fragility classification of AES features (SRQ2), and the impact of restricting to robust features on quality-prediction stability under AI assistance (SRQ3). Together, these results establish whether AI assistance produces a detectable, graded shift in AES features, which features drive that shift, and whether avoiding those features yields more stable quality assessment.

4.1 SRQ1: Detection Accuracy Across Intervention Types

Table 1 reports AUC-ROC and macro-F1 for the XGBoost and Random Forest classifiers, averaged over five GroupKFold folds, for each intervention level. Both metrics increase monotonically from Light to Heavy intervention, confirming that AI assistance creates progressively larger and more detectable distributional shifts in AES features.

Table 1: Binary classifier performance (original vs. AI-assisted essays) per intervention level. Metrics are mean ± SD across five GroupKFold folds (grouped by `essay_id` to prevent identity leakage). The random-guess baseline is AUC = 0.50, macro-F1 for a balanced 50/50 split is 0.50. Level 1 = grammar correction, Level 2 = style enhancement, Level 3 = substantive revision.

Level	Model	AUC-ROC	Macro-F1
1 (Light)	XGBoost	0.719 ± 0.003	0.655 ± 0.002
	Random Forest	0.695 ± 0.003	0.637 ± 0.004
2 (Medium)	XGBoost	0.921 ± 0.003	0.839 ± 0.003
	Random Forest	0.912 ± 0.003	0.830 ± 0.003
3 (Heavy)	XGBoost	0.999 ± 0.000	0.990 ± 0.001
	Random Forest	0.999 ± 0.000	0.986 ± 0.000

At Level 1 (grammar correction), XGBoost reaches AUC = 0.719 and macro-F1 = 0.655, well above the random-guess baseline (AUC = 0.50) but indicating that surface-only proofreading leaves many essays difficult to distinguish from their originals. At Level 2 (style enhancement), AUC rises to 0.921 and macro-F1 to 0.839, an increase of 0.20 and 0.18 points respectively over Level 1. Level 3 (substantive revision) produces near-ceiling detection: AUC = 0.999 and macro-F1 = 0.990, leaving almost no remaining gap to perfect separability. Random Forest tracks XGBoost closely at each level (within 0.02 AUC), confirming that the pattern is not model-specific.

4.2 SRQ2: Feature Fragility Analysis

Feature importance is quantified using global mean |SHAP| values averaged across all five folds and all three intervention levels. Table 2 lists all 26 features with their per-level and overall mean SHAP values and their fragility label. Features whose overall mean SHAP value falls at or above the third quartile (Q_3) of the distribution across all 26 features are labelled *fragile*, those at or below the first quartile (Q_1) are labelled *robust*, and the remaining features are *neutral*.

Seven features are labelled fragile. `avg_word_len` dominates: its mean SHAP value of 2.055 is 4.2 times larger than the next-highest feature (`pos_other_ratio`, 0.489), and its importance grows significantly across levels (0.585 → 1.423 → 4.156), indicating that word-length distributions are progressively pushed further from the human-writing baseline as intervention depth increases. Vocabulary richness (`mtld`, `dale_chall`) and discourse-level coherence (`avg_lexical_overlap`) also emerge as fragile, with their SHAP values increasing most strongly at Level 3, consistent with the Level 3 prompt explicitly targeting argumentation and paragraph organisation.

Seven features are labelled robust.
`word_count` and `sentence_count` are the two features that are most strongly correlated with human quality scores ($r = 0.67$ and $r = 0.61$ respectively on original essays; see Figure 7). They show small SHAP importance across all levels, indicating that AI assistance at all three intensities does not substantially alter essay length. The remaining robust features (*passive_ratio*, *pos_verb_ratio*, *pos_adv_ratio*, *mean_noun_phrase_modifiers*, *flesch_kincaid_grade*)

Table 2: Global mean |SHAP| importance per feature, reported separately for each intervention level and as an overall mean across levels. Features are sorted by overall mean SHAP (descending). Fragility labels: F = fragile (overall mean SHAP at or above Q_3 of the 26-feature distribution), R = robust (at or below Q_1), N = neutral.

Feature	L1	L2	L3	Overall	Label
avg_word_len	0.585	1.423	4.156	2.055	**F**
pos_other_ratio	0.425	0.511	0.530	0.489	**F**
smog_index	0.149	0.600	0.851	0.534	**F**
gunning_fog	0.100	0.127	1.347	0.525	**F**
mtld	0.022	0.413	0.909	0.448	**F**
dale_chall_readability	0.417	0.086	0.511	0.338	**F**
avg_lexical_overlap	0.146	0.152	0.709	0.336	**F**
avg_sent_len	0.328	0.214	0.457	0.333	**N**
automated_readability	0.067	0.357	0.408	0.277	**N**
pronoun_density	0.211	0.136	0.240	0.195	**N**
coleman_liau_index	0.270	0.195	0.107	0.190	**N**
mattr	0.056	0.293	0.218	0.189	**N**
hdd	0.054	0.169	0.312	0.178	**N**
connective_freq	0.022	0.198	0.254	0.158	**N**
pos_noun_ratio	0.098	0.139	0.134	0.124	**N**
pos_adj_ratio	0.013	0.171	0.179	0.121	**N**
linsear_write_formula	0.057	0.126	0.098	0.094	**N**
mean_dep_tree_depth	0.042	0.115	0.173	0.110	**N**
subordinate_clause_ratio	0.070	0.119	0.053	0.080	**N**
word_count	0.013	0.099	0.119	0.077	**R**
flesch_kincaid_grade	0.098	0.025	0.109	0.077	**R**
pos_verb_ratio	0.038	0.091	0.093	0.074	**R**
mean_noun_phrase_mod	0.055	0.093	0.072	0.073	**R**
pos_adv_ratio	0.011	0.047	0.056	0.038	**R**
passive_ratio	0.028	0.071	0.074	0.057	**R**
sentence_count	0.057	0.013	0.046	0.039	**R**

are syntactic and structural properties that the intervention prompts do not explicitly target.

The importance rankings are highly stable across cross-validation folds: all pairwise Spearman rank correlations between folds exceed $r = 0.95$ ($p < 10^{-13}$) at every intervention level, with a minimum of $r = 0.951$ (Level 3, folds 1 and 2) and a maximum of $r = 0.998$ (Level 2, folds 1 and 2). This confirms that the fragile/robust categorisation is not an artefact of any single fold's training data.

4.3 SRQ3: Quality Assessment with Robust Features

Table 3 reports Quadratic Weighted Kappa (QWK) for three feature conditions—all features, robust only, and fragile only—evaluated at each intervention level. Models are trained exclusively on original essays and applied to AI-assisted variants without retraining, so any QWK change reflects degradation caused by the distributional shift introduced by the intervention. Figure 2 shows the resulting trajectories, and Figure 3 summarises the absolute and relative drop from original to heavy intervention for each feature subset.

The all-features model achieves the highest QWK on original essays (0.731), briefly exceeding the 0.70 threshold, but degrades to 0.588 at heavy intervention, a relative drop of 19.5% (Table 3).

Table 3: Quality prediction performance (QWK) per feature subset and intervention level. Models are XGBRegressor instances trained on original essays (level 0) and evaluated on AI-assisted essays at each level without retraining. Values are mean ± SD across five GroupKFold folds. The dashed acceptability threshold in Figure 2 is QWK ≥ 0.70, following [14]. Relative drop is computed from level 0 to level 3.

Level	All	Robust	Fragile
Original	0.731 ± 0.010	0.682 ± 0.009	0.353 ± 0.011
Light	0.765 ± 0.001	0.709 ± 0.001	0.359 ± 0.004
Medium	0.709 ± 0.005	0.687 ± 0.002	0.233 ± 0.009
Heavy	0.588 ± 0.018	0.617 ± 0.004	−0.028 ± 0.031
Rel. drop	−19.5%	−9.6%	−107.9%

The robust-feature model begins lower on original essays (0.682) and does not reach 0.70 at any level, but degrades to only 0.617 at heavy intervention, a relative drop of 9.6% (Table 3, Figure 3).

The fragile-only model starts at 0.353 on original essays and falls to −0.028 at heavy intervention, a relative drop exceeding 100%, indicating predictions become systematically inverted relative to human scores at heavy intervention.

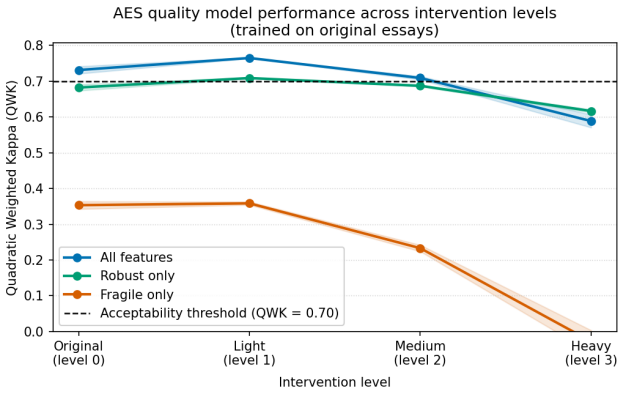


Figure 2: QWK by intervention level for all-features (blue), robust-only (green), and fragile-only (orange) models. Shaded bands show ±1 SD across five GroupKFold folds. The dashed line marks the acceptability threshold (QWK ≥ 0.70) following [14]. Neither the all-features nor the robust-only model consistently reaches this threshold, but the robust model degrades by only 9.6% from original to heavy intervention, compared with 19.5% for the all-features model, and remains above the all-features model from medium intervention onward. The fragile model collapses to a negative QWK at heavy intervention.

Figure 2 shows that the crossing point between the all-features and robust models occurs between medium and heavy intervention, after which the robust model maintains a QWK advantage of 0.029 points. This pattern is broadly consistent across the seven prompts (Appendix E), though absolute QWK varies considerably by prompt, ranging from 0.44–0.39 (robust, *A Cowboy Who Rode*

the Waves) to 0.72–0.68 (robust, *Facial action coding system*) across original to heavy intervention. Both length features (`word_count`, `sentence_count`) contribute substantially to the robust model's absolute performance. Their role is examined as a robustness check in Section 5.

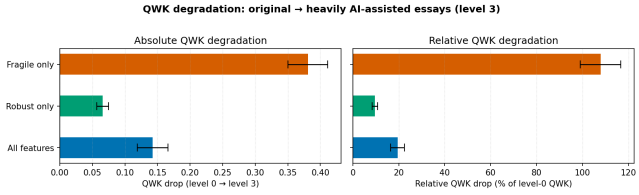


Figure 3: Absolute (left) and relative (right) QWK drop from original to heavy intervention (level 0 → level 3), per feature subset. Error bars show ±1 SD across five GroupKFold folds. The robust-feature model's absolute drop (0.066) falls below the 0.10 QWK criterion specified for SRQ3, while the all-features model's drop (0.143) exceeds it. The fragile-only model's drop (0.381) reflects its collapse to negative QWK at heavy intervention.

5 Discussion

5.1 Comparison with Prior Work

The detection results (SRQ1) are consistent with prior findings that lightly-edited AI text is harder to detect than fully AI-generated text [13]. At Level 1, XGBoost achieves AUC = 0.719, well above chance but far below the near-perfect detection reported for fully AI-generated text [4, 8]. At Level 3, AUC reaches 0.999, essentially matching full-generation detectability. This monotonic increase (Table 1) confirms the hypothesised intensity gradient. Prior detection work has largely treated AI involvement as a binary category [3, 4]. These results suggest the same approach extends naturally to a setting where AI involvement varies in degree, from light edits to substantial rewriting.

The SRQ2 finding that `avg_word_len` dominates feature importance (mean SHAP = 2.055, over four times the next-highest feature) aligns with prior observations that AI writing assistants systematically alter vocabulary and word-length distributions [15]. The clustering of readability indices among the fragile features is consistent with AI assistance, particularly at Level 2, directly targeting the surface properties these indices aggregate. This largely confirms the hypothesised ordering of surface-level and readability features as more fragile than discourse and syntactic features, with one exception: `flesch_kincaid_grade`, also a readability index, falls in the robust set rather than the fragile set. The high cross-fold stability of the fragility rankings (Spearman $r > 0.95$) suggests this categorisation reflects a genuine property of the feature space rather than a fold artefact, addressing a construct-validity concern raised in Section 3.5.

The SRQ3 results offer the clearest comparison with the broader AES evaluation literature. Williamson et al. [14] propose QWK ≥ 0.70 as the conventional acceptability threshold. The all-features model exceeds this at the original, light, and medium levels (0.731, 0.765, 0.709), falling below it only at heavy intervention (0.588). The robust-feature model sits just below threshold on original essays (0.682) and at medium intervention (0.687), briefly exceeds it at light intervention (0.709), and falls further below at heavy intervention (0.617). Both models are therefore threshold-adjacent under light-to-moderate intervention but fail under heavy intervention. However, SRQ3 was framed not only in terms of absolute acceptability but *relative degradation*: the robust model's absolute QWK drop (0.066) falls below the 0.10 SRQ3 criterion, while the all-features drop (0.143) exceeds it by more than a factor of two (Figure 3). The 0.029-point advantage of the robust-feature model at heavy intervention is large relative to the fold-level variability of both models (SD = 0.004 for the robust model, 0.018 for the all-features model), suggesting this crossover reflects a genuine difference rather than cross-validation noise. Restricting to robust features therefore substantially improves the *stability* of quality assessment under AI assistance, even though both models perform acceptably under light-to-moderate intervention. This is a more modest claim than the original hypothesis, but consistent with the broader pattern that feature-based AES models trained on human-written text show some degradation, though not necessarily collapse, when applied to out-of-distribution essays [1].

To my knowledge, no prior work reports QWK degradation for AES models under graded AI-assistance interventions specifically, so a direct effect-size comparison is not available. The closest comparator is work on AES robustness under domain shift more broadly [5], which similarly finds that feature-based models trained on one distribution of essays transfer imperfectly to another. The degradation observed here (9.6% to 19.5% relative QWK drop, depending on feature subset) is consistent with that general pattern, though not directly comparable in scale.

5.2 Why the Robust-Feature Model Still Falls Short of 0.70

A natural question is why the robust-feature model does not reach the 0.70 threshold even on original, human-written essays, where no distributional shift has occurred. Two non-exclusive explanations are plausible. First, by construction, the robust feature subset excludes the seven features most discriminative for detecting AI assistance, which are not necessarily the seven features least predictive of human quality scores. Figure 7 shows that several fragile features (e.g. `avg_word_len`, $r = 0.25$; `smog_index`, $r = 0.16$) carry non-trivial correlation with human scores on original essays. Removing them sacrifices some predictive signal even before any AI intervention is applied, which explains the 0.049 gap between the all-features and robust-feature models on original essays (Table 3). Second, the operationalisation of fragility in this thesis is based on classifier discriminativeness (SHAP importance for detecting AI assistance), not on predictive validity for human quality scores. These two properties are correlated but not identical, a feature can be highly stable under AI assistance (robust) while still being only weakly predictive of quality, or vice versa. The robust-feature set is therefore optimised for stability under intervention, not for maximal predictive power, and the SRQ3 results should be read as evidence about the former rather than the latter.

5.3 The Role of Length Features: An Ablation Check

Two of the seven features categorised as robust, `word_count` and `sentence_count`, are length proxies rather than substantive indicators of writing quality. Their robustness is, in one sense, unsurprising: Table 2 shows that AI assistance at all three intervention levels does not substantially alter these features' SHAP importance, consistent with the dataset-level observation (Section 3.1) that AI-assisted essays do not differ dramatically in length from their originals. At the same time, `word_count` and `sentence_count` are the two features most strongly correlated with human quality scores on original essays ($r = 0.67$ and $r = 0.61$ respectively; Figure 7), and permutation importance analysis of the robust-feature quality model (Figure 4) shows that these two features alone dominate the model's predictive power, with `word_count` contributing roughly four times the permutation importance of the next-ranked feature (`flesch_kincaid_grade`).

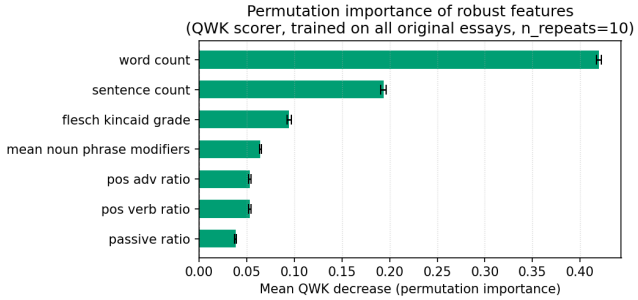


Figure 4: Permutation importance (mean QWK decrease, $n\_repeats = 10$) of the seven robust features in the quality model trained on original essays. `word_count` and `sentence_count` dominate, together accounting for the majority of the model's predictive power, while the remaining five features contribute substantially less.

To verify that the relative-robustness finding in Section 4 is not solely attributable to these two length proxies, I retrained all three quality models (all-features, robust-only, fragile-only) excluding `word_count` and `sentence_count`, leaving 24 features in total and five features in the robust subset (*flesch_kincaid_grade*, *pos_verb_ratio*, *pos_adv_ratio*, *passive_ratio*, *mean_noun_phrase_modifiers*). Figure 5 shows the resulting trajectories.

Without `word_count` and `sentence_count`, absolute QWK collapses across all three feature subsets. The all-features model falls from 0.731 to 0.523 on original essays, a drop of 0.208 from its 26-feature counterpart, and degrades further to 0.142 at heavy intervention, an absolute drop of 0.382 (72.9% relative). The robust-only model falls from 0.682 to 0.384 on original essays, a drop of 0.298, and reaches 0.263 at heavy intervention, an absolute drop of 0.121 (31.4% relative). The fragile-only model, already non-functional in the 26-feature analysis, remains so ($0.282 \rightarrow -0.035$).

The directional finding from Section 4 persists: the robust-only model's absolute drop (0.121) remains smaller than the all-features model's (0.382), and its relative drop (31.4%) remains less than

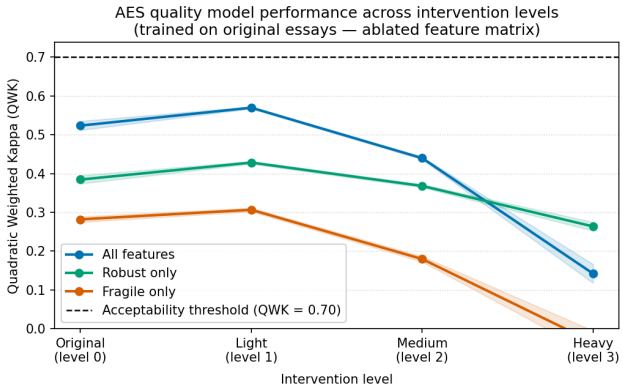


Figure 5: QWK by intervention level for all-features (blue), robust only (green), and fragile-only (orange) models, retrained without `word_count` and `sentence_count` (24 features total, 5 robust). The dashed line marks the 0.70 acceptability threshold [14]. All three models perform substantially worse in absolute terms than their 26-feature counterparts (Figure 2), but the robust-only model still degrades less than the all-features model from original to heavy intervention (0.121 vs. 0.382 absolute drop).

half that of the all-features model (72.9%). This indicates that the relative-robustness pattern is not solely an artefact of `word_count` and `sentence_count`. The five remaining robust features, while individually weak predictors, preserve a similar degradation profile.

However, the magnitude of the absolute collapse substantially qualifies the SRQ3 finding. Both the 0.121 absolute drop and the 31.4% relative drop for the length-excluded robust model exceed the 0.10 / 0.70 criteria specified for SRQ3 by a wide margin, and an original-essay QWK of 0.384 is far below any threshold that would support practical use. `word_count` and `sentence_count` therefore appear to be *necessary*, not merely convenient, for the robust-feature model's absolute performance to be meaningful at all: removing them costs 0.298 QWK points on original essays alone, more than four times the entire degradation the 26-feature robust model exhibits under heavy AI intervention (0.066).

This tension is best summarised as follows. The *relative*-robustness claim, that features which are stable under AI-assistance detection also yield quality models that degrade less under AI assistance, holds with or without length features. The *absolute*-performance claim, that a robust-feature model can support meaningful quality assessment, depends heavily on retaining `word_count` and `sentence_count` specifically. Given that word count alone explains approximately 50% of score variance in this dataset ($R^2 = 0.50$, Section 3.1), this is perhaps unsurprising, but it means the practical takeaway of SRQ3 is narrower than originally framed: it is not "non-discriminative features in general" that support robust assessment, but specifically essay length, which happens to be both highly predictive of quality and stable under the AI interventions studied here, alongside a smaller stabilising contribution from the remaining robust features.

5.4 Limitations

Construct validity. The central operationalisation of this thesis, equating feature fragility with classifier-assigned SHAP importance for AI-assistance detection, is valid only insofar as the classifier's decision boundary reflects genuine, intervention-induced feature shifts rather than confounds. The essay-level GroupKFold split (Section 3.5) guards against identity leakage, and the high cross-fold stability of the SHAP rankings ($r > 0.95$) suggests the categorisation is not a fold-specific artefact. However, fragility as defined here is a property of *detectability*, not directly of *quality-predictiveness*, and Section 5.2 shows these two properties are related but distinct. A feature could in principle be both highly detectable and highly predictive of quality (as appears to be partly the case for `avg_word_len`), in which case "fragile" and "quality-relevant" are not mutually exclusive categories, and removing fragile features necessarily trades some predictive power for stability.

Reliability. The fragile/robust/neutral categorisation depends on a quartile-based cutoff applied to each feature's global mean SHAP importance (itself averaged across cross-validation folds and the three intervention levels). While this cutoff is data-driven and produces a stable ranking across cross-validation folds, the choice of quartiles (rather than, say, the top or bottom six, eight, or ten features) is itself a methodological decision that was not varied systematically in this thesis. A threshold sensitivity analysis, varying the cutoff across a range of plausible values and checking whether the downstream SRQ3 pattern (smaller relative degradation for the robust subset) is preserved, would strengthen confidence that the qualitative conclusions are not an artefact of this specific cutoff. This is identified as a priority for future work.

Generalizability. The ASAP 2.0 dataset [2] consists of secondary-level essays from seven prompts, scored by a single human rater on a six-point scale, written before the widespread availability of LLM-based writing tools. External validity to other populations is limited in at least three ways. First, results may not generalise to university-level academic writing, which differs in genre conventions, baseline linguistic complexity, and the kinds of AI assistance students are likely to use (e.g. literature synthesis or argument construction, rather than the grammar-to-substantive-revision gradient simulated here). Second, the AI interventions were generated by a single model (GPT-4o-mini) under fixed decoding parameters. Different models, or models prompted differently, may produce edits with different feature signatures, and the fragility categorisation derived here is specific to this intervention protocol. Third, the dataset is English-only. Prior work has shown that AI-text detection performance and fairness can differ substantially across languages and for non-native writers [6], and the same may hold for feature fragility under AI assistance.

Scalability. The feature set used here (26 features across four families) was deliberately kept to established, computationally lightweight AES features, excluding language-model-based features such as perplexity and burstiness that were considered in early planning but removed to keep the feature space interpretable and the SHAP analysis tractable. This means the fragility findings apply to the kinds of features traditionally used in feature-based AES systems [7], but not to the neural, embedding-based representations used in more recent AES approaches [12]. Whether the fragile/robust distinction identified here transfers to neural AES systems, where "features" are learned representations rather than hand-engineered statistics, is an open question.

Reproducibility. All feature extraction, classification, and quality-prediction code, along with the AI intervention prompts, will be made publicly available. Random seeds are fixed throughout, and the GPT-4o-mini intervention generation uses temperature 0 with a fixed seed of 42. One caveat to full reproducibility is that API-based generation is not guaranteed to be perfectly deterministic across time even with fixed parameters, so exact reproduction of the AI-assisted essay variants is not guaranteed, although the feature-level and classifier-level findings are expected to be robust to minor variation in the generated text, given that the SHAP rankings are already stable across five independently-trained folds.

5.5 Alternative Interpretations and Scientific Value

An alternative reading of the SRQ3 results is that the robust-feature model's smaller relative degradation is not evidence of genuine feature-level robustness, but simply a consequence of the robust-feature model having less room to fall: its original-essay QWK (0.682) is already closer to the fragile-only model's collapse point than the all-features model's is. Under this reading, a model with lower baseline performance will, almost mechanically, show a smaller relative drop, since relative drop is normalised by a smaller starting value, while an absolute drop of similar magnitude would appear proportionally larger when starting from 0.353 (the fragile-only baseline) than from 0.731. Examining the absolute drops directly (Figure 3) partially addresses this: the robust-feature model's absolute drop (0.066) is also smaller than the all-features model's (0.143), not just proportionally but in raw QWK points, which is harder to explain purely as a baseline artefact. Nonetheless, this tension between absolute and relative framings of "robustness" is worth making explicit, and future work could report both consistently rather than emphasising one.

Despite the qualifications above, this thesis makes a methodological contribution that the results, even where they fall short of the original hypotheses, help to substantiate: the *detection as diagnosis* framework, using a classifier's discriminative power for AI-assistance detection as a diagnostic lens on AES feature stability, produces categorisations (fragile vs. robust) that are stable across cross-validation folds and that correspond to a measurable, if modest, difference in downstream quality-prediction stability. This bridges the AES and AI-detection literatures in the manner motivated in Section 1: features that are diagnostic of AI assistance are, at least in this dataset, also the features whose removal yields more stable (though not necessarily more accurate) quality assessment under AI assistance. The research gap, the disconnect between AES research (which assumes human authorship) and AI-detection research (which treats generation as a binary outcome), is addressed by showing that these two perspectives can be operationally linked through feature importance, even if the practical payoff (a deployment-ready, AI-robust AES model) remains only partially realised.

6 Conclusion

This thesis addressed a gap at the intersection of automated essay scoring (AES) and AI-text detection: while AES research assumes human authorship and detection research treats AI involvement as a binary outcome, little was known about whether AES features remain reliable under AI-assisted writing, or how different intensities of AI assistance affect different feature categories. To address this, I proposed a *detection as diagnosis* framework, using a classifier's ability to discriminate AI-assisted from human-written essays as a diagnostic lens on AES feature fragility.

The three sub-questions were answered as follows. Detection accuracy (SRQ1) increases monotonically with intervention intensity, from AUC = 0.719 under grammar correction to AUC = 0.999 under substantive revision, confirming a genuine intensity gradient. Seven of 26 features were identified as fragile (SRQ2), dominated by `avg_word_len`, and seven as robust, with stable categorisations across folds ($r > 0.95$). A quality model restricted to robust features (SRQ3) did not consistently exceed the conventional QWK ≥ 0.70 threshold [14], but degraded substantially less under heavy intervention (absolute drop 0.066, within the 0.10 SRQ3 criterion) than a model using all features (drop 0.143).

Taken together, these results answer the main research question: AES features can distinguish AI-assisted from human-written text with increasing reliability as intervention intensity grows, and a subset of features that are non-discriminative for this detection task does support more stable, though not necessarily more accurate, quality assessment under AI assistance. Robust quality assessment under AI assistance therefore appears feasible to a meaningful but limited extent: stability, not absolute deployment-readiness, is what non-discriminative features deliver.

This conclusion is qualified in two ways (Sections 5.2, 5.3): the robust-feature model's absolute QWK remains below 0.70 even on unassisted essays, and the stability advantage depends specifically on retaining the length-based features `word_count` and `sentence_count`, without which absolute QWK collapses (0.682 $\rightarrow$ 0.384). The main conclusion therefore narrows to: robust features, and especially essay length, support more stable but not deployment-ready AES under AI assistance, for this dataset and intervention protocol.

This research nonetheless adds value relative to the state of the art by demonstrating that detection-based feature importance can be operationally linked to AES quality-model stability, bridging two literatures that have so far developed largely independently, and by providing a dataset-grounded characterisation of which feature categories are disrupted by which intervention intensities. The results are specific to ASAP 2.0, a single AI model (GPT-4o-mini) under one intervention protocol, and a fixed 26-feature set of established, hand-engineered AES features.

Future work should prioritise three directions. The most immediate is the threshold sensitivity analysis identified in Section 5.4, varying the fragility cutoff across a range of plausible values to establish whether the relative-robustness finding, and the specific role of length features identified in the ablation, hold beyond the quartile-based cutoff used here. A second direction is to repeat the intervention and detection pipeline with a different AI model, to test whether the fragility categorisation reflects properties of GPT-4o-mini's editing behaviour specifically or generalises across writing assistants more broadly. A third, longer-term direction is to extend this framework beyond hand-engineered features to the embedding-based representations used in neural AES systems, where it remains an open question whether an analogous notion of "fragility under AI assistance" can be defined and whether it would similarly concentrate in a small subset of learned dimensions.

References

[1] Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 503–509. doi:10.18653/v1/P18-2080

[2] Scott A. Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A Large-Scale Corpus for Assessing Source-Based Writing Quality: ASAP 2.0. *Assessing Writing* 65 (2025), 100954. Dataset available from https://www.kaggle.com/datasets/lburleigh/asap-2-0.

[3] Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12463–12492.

[4] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. doi:10.48550/ARXIV.1906.04043

[5] Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art.. In *IJCAI*, Vol. 19. 6300–6308.

[6] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023).

[7] Danielle S. McNamara, Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. Coh-Metrix: Capturing Linguistic Features of Cohesion. *Discourse Processes* 47, 4 (May 2010), 292–330. doi:10.1080/01638530902959943

[8] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*. PMLR, 24950–24962.

[9] OpenAI. 2023. New AI classifier for indicating AI-written text. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

[10] Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing* 20 (April 2014), 53–76. doi:10.1016/j.asw.2013.04.001

[11] Miriam Sullivan, Andrew Kelly, and Paul McLaughlan. 2023. ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching* 6, 1 (March 2023). doi:10.37074/jalt.2023.6.1.17

[12] Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1882–1891.

[13] Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity* 19, 1 (2023), 1–39.

[14] David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice* 31, 1 (2012), 2–13. doi:10.1111/j.1745-3992.2011.00223.x

[15] Da Yan. 2023. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and information technologies* 28, 11 (2023), 13943–13967.

Appendix A Feature SHAP Importance (26-Feature Set)

Figure 6 shows global mean |SHAP| importance for the full 26-feature set described in Section 4 (Table 2).

Appendix B Extracted Features

Table 4 lists the full set of 26 features used in the classification and quality-prediction pipelines (Section 3), grouped by family.

Table 4: Features extracted, grouped by family.

Feature	Description
Surface	
word_count	Total number of words
sentence_count	Total number of sentences
avg_sent_len	Average number of words per sentence
avg_word_len	Average number of characters per word
mattr	Lexical diversity via sliding window of 50 words
mtld	Length-invariant lexical diversity measure
hdd	Hypergeometric lexical diversity, most robust to text length
pos_noun_ratio	Proportion of tokens that are nouns
pos_verb_ratio	Proportion of tokens that are verbs
pos_adj_ratio	Proportion of tokens that are adjectives
pos_adv_ratio	Proportion of tokens that are adverbs
pos_other_ratio	Proportion of tokens that are none of the above
Readability	
flesch_kincaid_grade	Grade level based on syllables and sentence length
coleman_liau_index	Grade level based on characters and sentences per 100 words
gunning_fog	Grade level penalising words with 3+ syllables
smog_index	Grade level based purely on polysyllabic word counts
automated_readability_index	Grade level based on character and word counts
dale_chall_readability_score	Difficulty score penalising words outside a 3000-word familiar-word list
linsear_write_formula	Grade level based on ratio of easy to hard words
Coherence	
connective_freq	Number of discourse connectives per sentence
avg_lexical_overlap	Mean Jaccard similarity of content lemmas between adjacent sentences
Syntactic	
mean_dep_tree_depth	Average depth of the dependency parse tree per sentence
subordinate_clause_ratio	Number of subordinate clauses per sentence
passive_ratio	Proportion of sentences containing a passive construction
mean_noun_phrase_modifiers	Average number of modifiers per noun chunk
pronoun_density	Proportion of tokens that are pronouns

Appendix C Hyperparameter Settings

Table 5: Hyperparameter settings for all models. All models use random_state/random_seed = 42. Parameters not listed were left at library defaults.

Model	Parameter	Value
XGBClassifier (SRQ1, SRQ2)	n_estimators	300
	max_depth	6
	learning_rate	0.05
	scale_pos_weight	per-fold, $n_{neg}/\max(n_{pos}, 1)$
	eval_metric	auc
RandomForestClassifier (SRQ1, SRQ2)	n_estimators	300
	n_jobs	-1
XGBRegressor (SRQ3)	n_estimators	300
	max_depth	6
	learning_rate	0.1
	subsample	0.8
	colsample_bytree	0.8

Appendix D Feature-Score Correlation (Original Essays)

Figure 7 reports Pearson r between each feature and human essay scores on original essays, supporting the discussion in Sections 5.2 and 5.3.

Appendix E Per-Prompt Quality Prediction Results

Figure 8 reports Quadratic Weighted Kappa (QWK) for each of the seven ASAP 2.0 prompts individually, for the all-features, robust-only, and fragile-only models across all four intervention levels. This complements the aggregate results in Section 4 by showing that the relative robustness of the robust-feature model is not driven by any single prompt.

Appendix F Per-Prompt Quality Prediction Results (Ablated Feature Set)

Figure 9 reports per-prompt QWK for the all-features, robust-only, and fragile-only models retrained without `word_count` and `sentence_count` (24 features total, 5 robust; see Section 5.3). This complements Appendix E by showing that the directional pattern, smaller relative degradation for the robust-only model than the all-features model, persists across all seven prompts even when length features are excluded, despite the substantially lower absolute QWK values.

Appendix G Feature SHAP Importance (Ablated Feature Set)

Figure 10 shows global mean |SHAP| feature importance for the 24-feature ablated model (i.e. the 26 features used in Section 4 excluding `word_count` and `sentence_count`), with fragility labels recomputed using the same quartile-based procedure described in Section 4. The overall ranking is largely preserved, `avg_word_len` remains the dominant feature by a wide margin, but two features change category relative to the primary 26-feature analysis (Table 2): `dale_chall_readability_score` shifts from fragile to neutral, and `avg_sent_len` shifts from neutral to fragile. These shifts reflect the redistribution of relative importance once two high-correlation length features are removed, rather than a substantive change in which linguistic properties are most affected by AI assistance. The robust feature subset used in the ablation (Section 5.3) is defined relative to this 24-feature ranking.

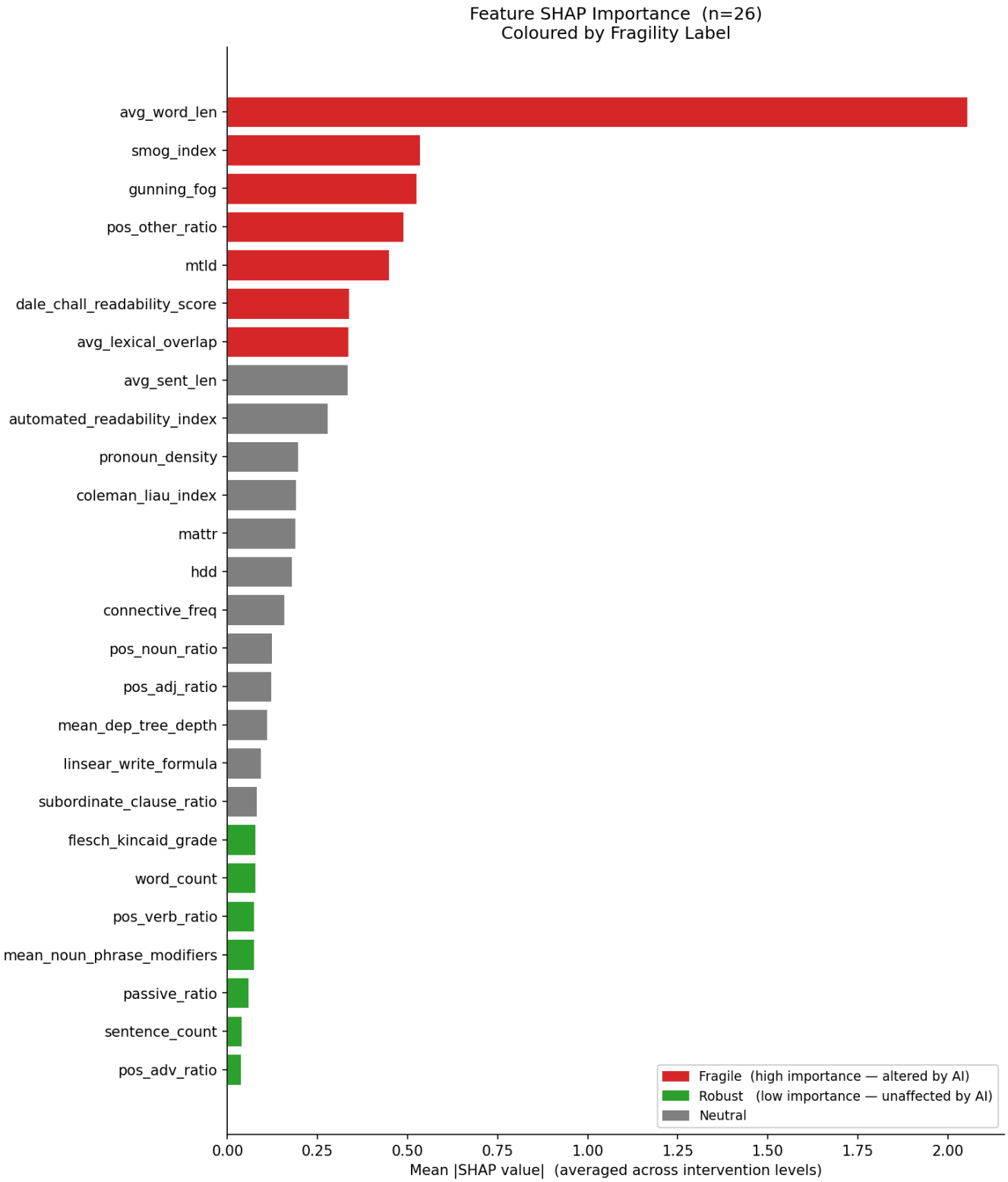


Figure 6: Global mean |SHAP| importance for all 26 features, averaged across intervention levels and cross-validation folds. Bars are coloured by fragility label: red = fragile (mean SHAP at or above Q_3 of the 26-feature distribution), green = robust (at or below Q_1), grey = neutral.

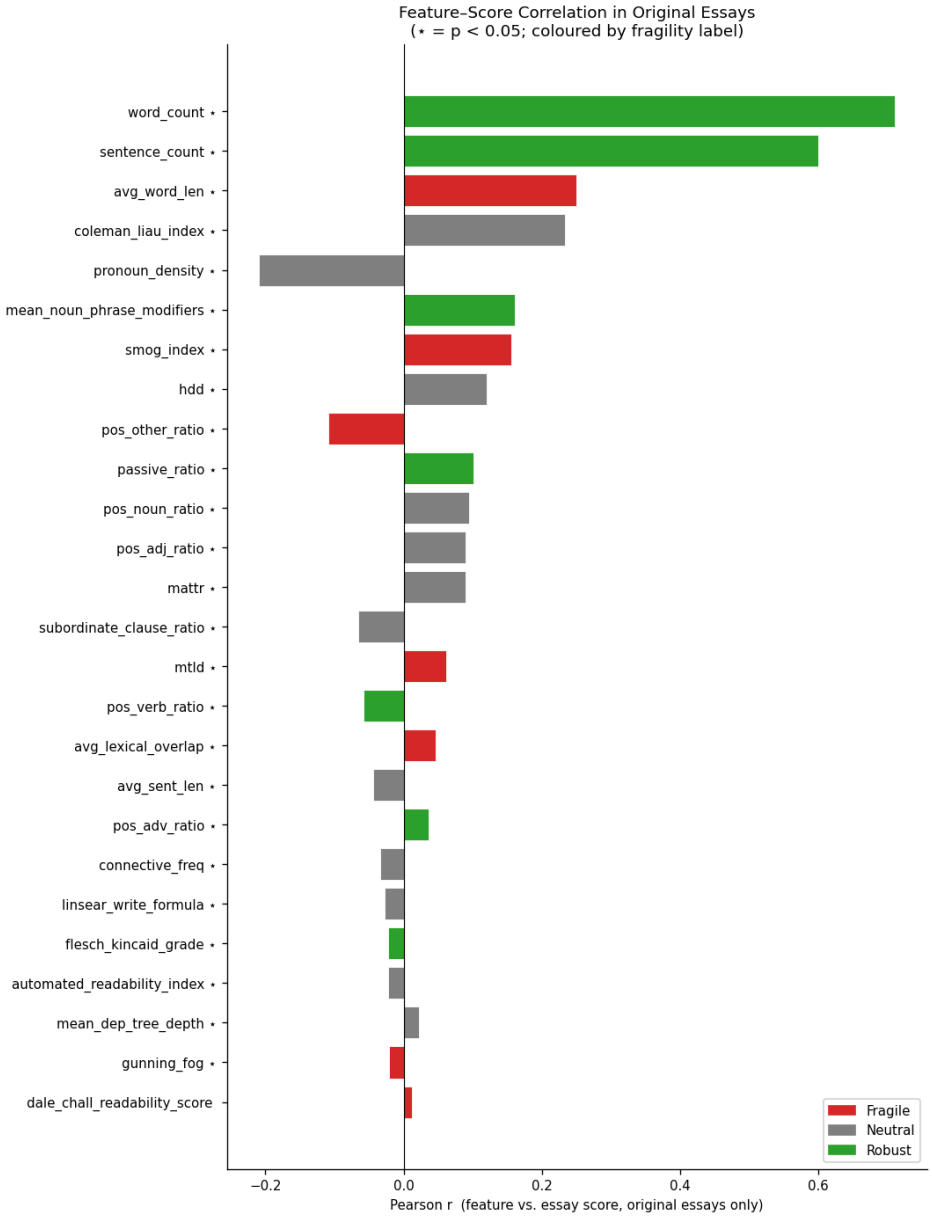


Figure 7: Pearson r between each feature and human essay scores on original essays only, coloured by fragility label. Asterisks mark $p < 0.05$, 25 of 26 features are significant at this threshold (`dale_chall_readability_score` is the sole exception, despite being classified as fragile). The two features most predictive of human quality (`word_count`, $r = 0.67$, `sentence_count`, $r = 0.61$) are both robust, suggesting that essay length is preserved by AI assistance and remains a valid quality signal. Several fragile features (e.g. `avg_word_len`, $r = 0.25$, `smog_index`, $r = 0.16$) show moderate positive correlations with quality, meaning their disruption by AI assistance directly undermines scoring validity.

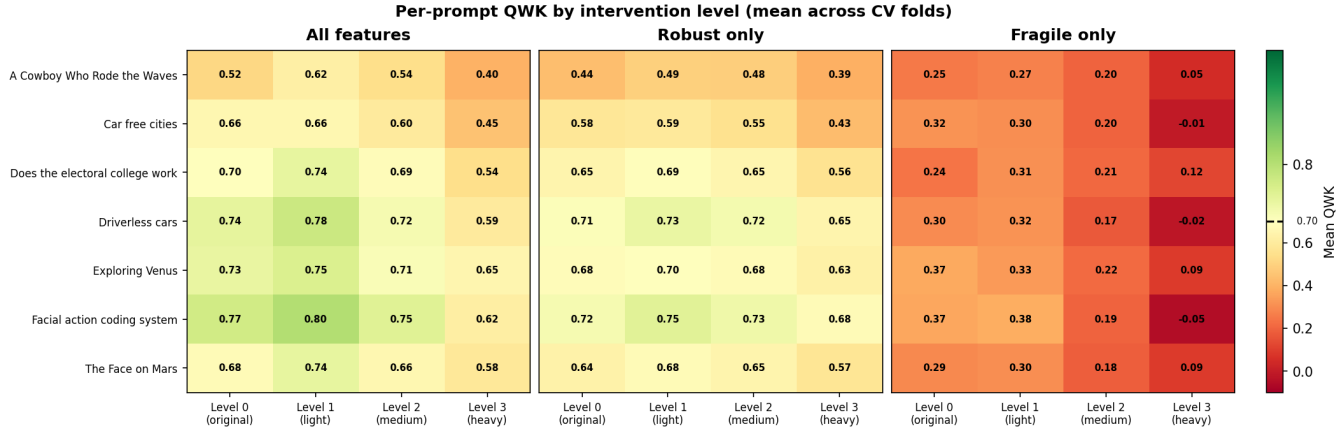


Figure 8: Per-prompt QWK by intervention level (mean across five GroupKFold folds), for all-features, robust-only, and fragile-only models (26-feature set, as in Section 4). Cell values are mean QWK; colour encodes the same scale across all three panels. The dashed line on the colour bar marks the 0.70 acceptability threshold [14]. *A Cowboy Who Rode the Waves* performs worst across all conditions, while *Facial action coding system* performs best. The pattern of smaller degradation for the robust-only model relative to the all-features model (Section 4) holds across all seven prompts. Section 5.3 shows this directional pattern persists, but at substantially lower absolute QWK, when the length-based robust features (`word_count`, `sentence_count`) are excluded.

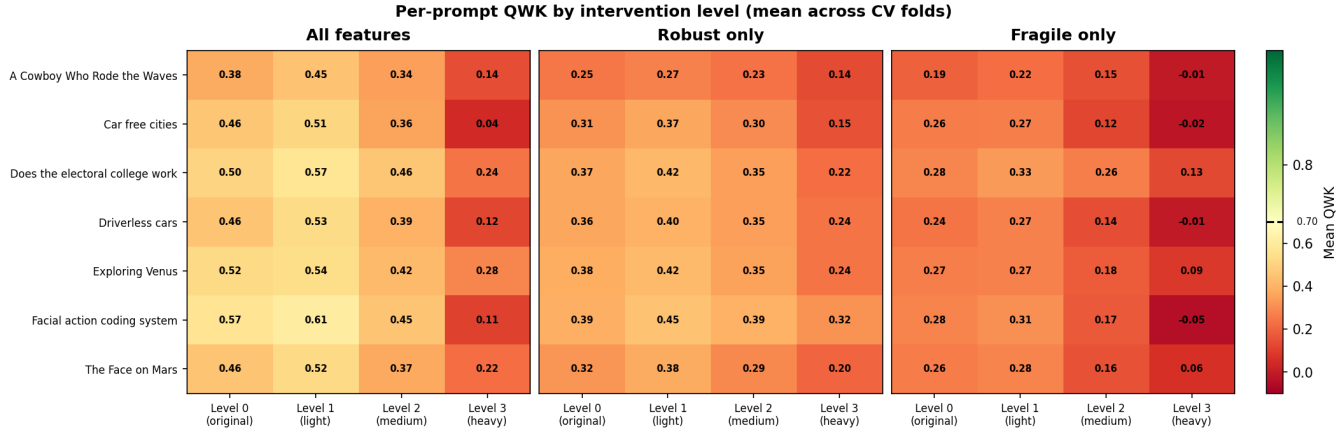


Figure 9: Per-prompt QWK by intervention level (mean across five GroupKFold folds), for all-features, robust-only, and fragile-only models, retrained without `word_count` and `sentence_count` (24 features total, 5 robust). Cell values are mean QWK; colour encodes the same scale across all three panels. The dashed line on the colour bar marks the 0.70 acceptability threshold [14]. Compared with the 26-feature results (Appendix E), absolute QWK is substantially lower across all prompts and conditions, but the robust-only model's smaller degradation relative to the all-features model holds across all seven prompts.

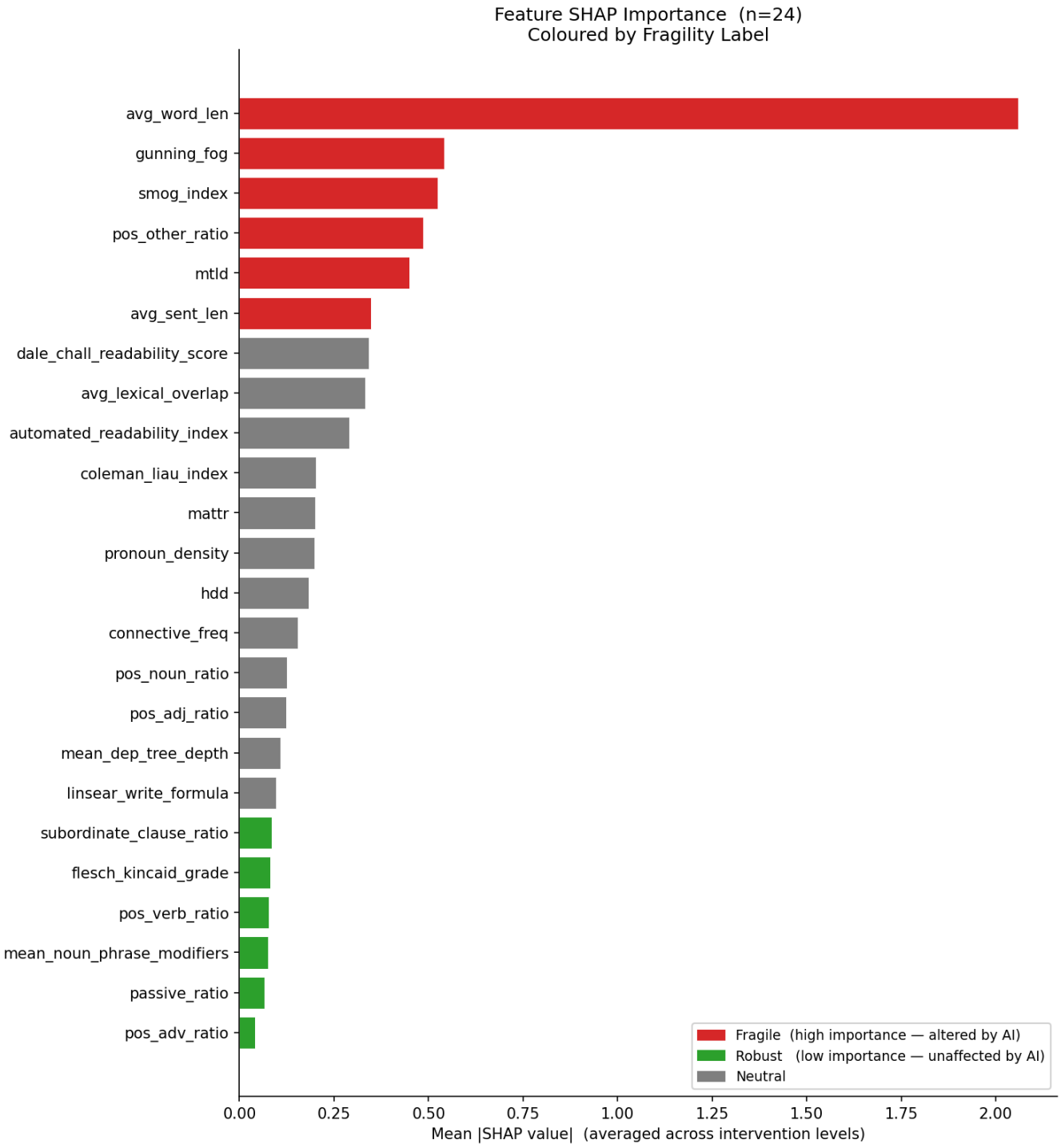


Figure 10: Global mean |SHAP| importance for the 24-feature ablated model (26 features excluding `word_count` and `sentence_count`), averaged across intervention levels and cross-validation folds. Bars are coloured by fragility label (recomputed for this feature set): red = fragile, green = robust, grey = neutral. `avg_word_len` remains dominant, consistent with the primary 26-feature analysis (Figure 6).