Proffiliadur: Welsh Language Text Profiling Toolkit

Nicolás Gutiérrez-Rolón[1], **Jonathan Davies**[3], **Tomos Williams**[1], **Dawn Knight**[2], **Fernando Alva-Manchego**[1]
[1]School of Computer Science and Informatics, Cardiff University, UK
[2]School of English, Communication and Philosophy, Cardiff University, UK
[3]School of Welsh, Cardiff University, UK
{GutierrezRolonN, DaviesJW9, KnightD5, AlvaManchegoF}@cardiff.ac.uk
tomos11@icloud.com

Abstract

We introduce Proffiliadur, a Python toolkit for text profiling and readability analysis in Welsh. The toolkit computes 141 surface, lexical, morphological, and syntactic indices, designed to capture linguistic variation while incorporating a Welsh-specific tokenisation process that enables accurate morphological analysis and handles phenomena such as initial consonant mutation. Proffiliadur enables systematic assessment of text accessibility and supports applications in education, healthcare, and public communication. We demonstrate the toolkit's usefulness through two complementary analyses. First, we examine texts written in accordance with the *Cymraeg Clîr* ("Clear Welsh") principles and compare them with regular Welsh texts. Second, we analyse texts across Common European Framework of Reference for Languages (CEFR) proficiency levels to explore how linguistic complexity varies with learner ability. We also evaluate feature-based and neural classification models for automatic complexity detection, showing that interpretable linguistic indices alone achieve strong predictive performance ($F_1 = 0.94$), comparable to a fine-tuned transformer ($F_1 = 0.97$). Proffiliadur provides the first dedicated text profiling toolkit for Welsh, offering reproducible, linguistically grounded measures of readability for a low-resource language.

Keywords: text profiling, linguistic features, readability, low-resource, Welsh

1. Introduction

Readability assessment measures how easily different audiences can understand a given text. It bridges NLP and the study of text complexity, providing a vehicle to evaluate a text's accessibility for readers with varying levels of language proficiency. Studies on official and health-related communication (Friedman et al., 2006; Walsh and Volsko, 2008; Serry et al., 2022) highlight that overly complex texts restrict engagement and understanding, especially among vulnerable groups. Therefore, ensuring that written materials are accessible is a crucial component of effective public communication. In this regard, readability metrics have become essential tools for ensuring accessible communication. They support the development and evaluation of educational resources, government and legal documents, and health communication materials, improving user experience and understanding across diverse audiences.

Despite a substantial body of work on readability assessment for English, research into metrics for minority and low-resource languages remains limited. Projects that have sought to develop a baseline model for Cebuano (Reyes et al., 2022), Tibetan (Wang et al., 2019) and Basque (Gonzalez Dios, 2016) have resulted in existing formulae that consider word and sentence length, but are of little use when applied to different languages that are more morphologically complex. Recent multilingual approaches using deep embeddings (Dalal et al., 2023) and deep-neural models (Imperial, 2021) show that, while promising, they still depend on large parallel datasets which are scarce in low-resource languages. Together, these studies highlight a methodological and resource gap: many minority/low-resource languages lack sufficient corpus data, computational tools and valid benchmarks with which to measure readability.

To address this need, we present **Proffiliadur**, a Python toolkit for text profiling and readability analysis in Welsh. The toolkit computes 141 surface, lexical, morphological, and syntactic indices, incorporating a Welsh-specific tokenisation process that enables accurate morphological analysis and handles phenomena such as initial consonant mutation. The toolkit allows users to extract interpretable linguistic features that can be applied to a variety of tasks, from readability prediction to stylistic analysis and corpus comparison.

We evaluate **Proffiliadur** through two complementary analyses and a classification experiment. First, we use the toolkit to compare texts written in line with *Cymraeg Clîr* principles against regular Welsh texts, identifying which linguistic features best characterise "clear" and "not clear" texts. Second, we analyse Welsh texts aligned with CEFR levels (A1–A2) to examine how linguistic complexity varies with language proficiency. Finally, we train and evaluate classification models (feature-based, embedding-based, and neural) to automatically distinguish between "clear" and "not clear" texts, thus testing the predictive power of the toolkit's indices

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1129–1142
11–16 May 2026.

against modern neural representations.

The results show that linguistic features derived from *Proffiliadur* alone achieve strong predictive performance ($F_1 = 0.94$), nearly matching a fine-tuned large language model ($F_1 = 0.97$) while offering more interpretability and lower computational cost. The analyses also provide empirical evidence supporting the *Cymraeg Clîr* guidelines, such as the use of impersonal tense, explicit subjects, shorter clauses, and preference for periphrastic constructions.

In sum, the main contributions of this work are:

- Proffiliadur, the first Welsh-specific toolkit for text profiling and readability analysis;[1]
- empirical insights into the stylistic and linguistic features associated with "clear" Welsh text and language proficiency; and
- baseline classification experiments comparing feature-based, embedding-based, and large language models for "clear" Welsh prediction.

2. Related Work

2.1. Text Profiling Tools

Readability indices are automatically calculated metrics that assess how easy a text can be read and understood at specific levels of comprehension, based on, for example, their lexical diversity and linguistic complexity.

The use of readability metrics has proven impactful across several major languages. For example, Kraus (2025) examined a web-based application, drawing on 50 different variables to improve the clarity of legal texts in Czech. Leal et al. (2024) developed readability tools for Brazilian Portuguese for use in education contexts. In English, readability tools have been used to characterise texts produced by individuals with dementia (Toledo et al., 2018), examine financial forecasts and reporting (Loughran and McDonald, 2015) and to assist in classifying learner texts according to CEFR levels (Uchida, 2025).

While a number of digital readability tools exist, including Text Inspector, Grammarly, MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021) and CohMetrix (McNamara et al., 2014), tools which support the analysis of minority and low-resource languages such as Welsh are often scarce.

2.2. The Welsh Language Context

Welsh is an official language of Wales. While deeply embedded in many aspects of daily life in Wales, including social and family settings, education, professional communication, literature, and the arts, Welsh is considered a minority language due to its low numbers of speakers, particularly when compared to the more dominant English language. Welsh is also considered a low-resource language because its digital and technological resources are less developed than those available for more dominant languages.

Efforts to close the resource gap for Welsh are well underway, and initiatives by organisations such as Canolfan Bedwyr and DigiGrid are making significant progress in this area. Government policies like *Cymraeg 2050* (Welsh Government, 2017) and the Digital Action Plan (Welsh Government, 2018) further support the development of Welsh language tools, aiming to integrate Welsh more fully into everyday life.

One major challenge in developing readability metrics for Welsh and other minority languages is the lack of large scale corpora. The creation of resources such as CorCenCC (Knight et al., 2020b), has begun to address this issue by providing valuable linguistic data that has been collected in a principled way. Yet, this corpus has not been analysed to produce metrics that assess text readability.

Beyond linguistic resources, Welsh presents unique linguistic challenges. Features such as short-form verbs, initial consonant mutations, and strong regional dialectic differences complicate the development of effective readability tools. Additionally, Welsh's sociolinguistic status adds further complexity to the available data. For example, (Cunliffe et al., 2022) sought to create a series of metrics that can be used to assess Welsh texts and found that it was necessary for tools to be able to distinguish between languages even within text, as code-switching in Welsh language communication is commonplace (p. 3).

Readability assessment could serve as an important bridge between Welsh Government language policy and everyday language use among speakers. By adopting the guidelines of Clear Welsh (*Cymraeg Clîr*) (Williams, 1999) as a baseline within a broader framework that incorporates diverse contextual sources, a readability tool could assess the lexical profile of texts against both official standards and everyday usage. Such metrics would be valuable not only in academic research but also in community settings. For instance, they could help public bodies simplify the language used in official communications with the public.

Although the *Cymraeg Clîr* guidance promotes plain Welsh (encouraging, for instance, shorter sentences and everyday vocabulary over traditional forms), it currently lacks computational means to assess compliance or evaluate readability at scale.

In addition to addressing resource gaps, there

[1] Code available at https://github.com/cardiffnlp/proffiliadur

is a pressing need for targeted tools in the Welsh language contexts. For example, the National Survey of Adult Skills (OECD, 2011) has highlighted the low levels of literacy within Wales, particularly in Welsh when compared to English in Wales. This underscores the importance of developing effective readability resources tailored to the needs of Welsh speakers; a gap **Proffiliadur** aims to fill.

3. The Proffiliadur Toolkit

Proffiliadur computes a broad range of indices representing different linguistic dimensions, including surface, lexical, morphological, and syntactic aspects. It leverages a part-of-speech (PoS) tagger, a dependency parser, and lexical resources for vocabulary profiling. The indices include both absolute counts and normalized frequencies, which are suitable for tasks such as text classification and readability assessment.

This section provides an overview of the toolkit, describing the types of indices extracted, the extraction pipeline, the external tools and resources employed, and the adaptations implemented to address Welsh-specific linguistic characteristics.

3.1. External Tools and Resources

The toolkit makes use of several external resources:

Vocabulary Lists. Three main types of lists are used for lexical profiling:

- Frequency-based lists (K1-K6+), in which each band represents a set of 1,000 words ranked by frequency, which were compiled as *Yr Amliadur* (Knight et al., 2020c) alongside the CorCenCC (Knight et al., 2020a) project;[2]
- WJEC/CBAC lists (A1-B1, representing CEFR levels) compiled from vocabulary lists provided by Welsh Joint Education Committee (WJEC) (675 words in A1, 652 in A2, 756 in B1); and
- Stopwords lists (437 and 486 words), with one list including Welsh-only words and the other also including English stopwords.

PoS Tagging. We rely on CyTag2 (Neale et al., 2018), a Welsh rule-based PoS tagger developed as part of the CorCenCC project. CyTag2 is used for tokenisation, as well as both coarse and fine-grained morphological annotation.

Dependency Parsing. A Universal Dependencies parser for Welsh from UDPipe (Straka and Straková, 2017) is used to extract syntactic structures for every sentence in the texts.

[2] https://www.corcencc.org

3.2. Linguistic and Stylistic Features

The feature set includes 141 indices spanning multiple levels of linguistic analysis, designed to capture lexical and morphosyntactic variation. Each feature is computed using different types of measures, including absolute counts, means, unique incidences, and ratios, in order to provide a detailed and comparable linguistic profile.

Absolute numbers correspond to raw counts of linguistic units, such as the number of verbs, nouns, or sentences. Unique incidence measures express unique occurrences. Ratio scores compare the relative frequency of one linguistic characteristic to the broader class it belongs to (e.g. ratio of feminine nouns to all nouns, or subordinate to all conjunctions). Mean and averages capture distributional tendencies, providing length-independent density measures.

The following groups summarise the major categories of extracted features:

3.2.1. Surface-level Indices

These 11 general indices are language agnostic (with tokenising considerations) and are essential for normalisation. They include raw text descriptors such as Token Count, as well as aggregated statistics such as Average Sentence Length (words per sentence). The TTR (Type-Token ratio) and TTR (Token-Type ratio) capture lexical diversity, while adapted readability formulae assess text complexity using Welsh-specific readability formulae (see Sec. 3.3) that integrate sentence length, word density, and interpunctual distance. These measures serve as baseline indicators and control variables for interpreting higher-level linguistic complexity.

3.2.2. Lexical Indices

This set includes 33 indices that measure the absolute and relative frequency of words in different bands according to frequency lists (see Sec. 3.1). For each band, the system computes (a) *unique matches* - the number of distinct lemmas at that level, (b) *total appearances* - the total token frequency across the text, and (c) *ratio to all words* - the proportion of tokens in that level relative to all tokens.

3.2.3. Morphological Indices

This set is the largest group with 77 indices derived using the PoS tagger CyTag. It includes coarse and fine-grained PoS counts, including total and average frequencies of nouns, verbs, adjectives, adverbs, prepositions, pronouns, articles, determiners, particles, numerals, conjunctions, and punctuations, as well as normalised frequencies of per-sentence averages.

Verb-related indices include number of verbs, unique verb lemmas, and ratios for each tense and mood: Present, Past, Imperfect, Future, Imperative, Conditional, Subjunctive, Future Subjunctive, and Impersonal. Ratios express the proportion of verbs in each tense relative to all verbs in the text. Verb noun counts and their ratios capture the periphrastic structures of Welsh.

Noun-based indices include number and ratios of Masculine, Feminine, Singular, Plural, and Proper nouns.

Other morphological indicators capture coordination and subordination with respect to conjunctions, and average distributions of modifiers like adjectives, adverbs, prepositions, numerals, determiners, and grammatical particles.

Finally, Welsh-specific morphological features include counts and averages of *Soft*, *Nasal*, and *Aspirate* mutations, as well as *h-prothesis* (known in Welsh as *Anadliad Caled*).

3.2.4. Syntactic Indices

This group includes 20 features extracted using UDPipe dependency parsing, and capture structural aspects of Syntax. Key features include *Max Tree Depth*, *Max Tree Breadth*, *Max Branching Factor*, and *Average Head Distance*, indicators of sentence complexity and dependency layering.

Subject and object analyses provide ratios and average depths for each, along with their average head distances, reflecting how far arguments are from verbs in their respective dependency structures. Indices like *Subject Ratio*, *Object Ratio*, and *Left-Right Branching Ratio* capture asymmetries in sentence organisation and word order.

Clause-level metrics include *Total Clause Count*, *Clause Count per sentence*, and *Clause Density (clauses per token)*, offering measures of subordination frequency. Other metrics such as *Number of Auxiliaries* and *Number of Coordinated Conjuncts* describe the degree of sentence embedding.

These syntactic measures collectively quantify structural, lexical and grammatical complexity.

Together, these 141 linguistic indices provide a comprehensive representation of textual structure and language use across surface, lexical, morphological, and syntactic levels.[3]

3.3. Welsh-specific Adaptations

Several parts of the indices computation pipeline require special considerations to the intricacies and nuances of the Welsh language.

Tokenisation. We use CyTag2 to both maintain consistency and ensure correct handling of Welsh orthography and grammar. This is crucial to properly process apostrophes in contractions and clitics, which can significantly affect word counts and, therefore, all features normalised by word count. For example, *'n* can carry several meanings. It is commonly used as a contraction of '*yn*', a clitic functioning as a predicative complement to verbs, or a marker meaning 'is' before adjectives. It also appears in contractions of conjugated verbs like (*roeddwn* and *roedden* → *ro'n*), and as an infixed possessive pronoun (our) alongside prepositions (*gyda'n teulu* "with our family").

Morphology. Morphological features from CyTag2 are tailored to Welsh-specific grammatical categories, many of which may not have direct equivalents in other languages. For instance, tenses and moods like the future subjunctive (*dibynnol dyfodol*) or impersonal mood (*amhersonnol*) are components of the indices that are Welsh-specific.

Readability Formulae. Three traditional readability formulae are included as indices in the toolkit, which were adapted for Welsh and other languages by (Matricciani, 2023) through statistical scaling methods applied to the Italian GULPEASE readability index (Lucisano and Piemontese, 1988; Matricciani, 2023), and by extending it to include a variable of interpunctual distance.

Mutations. Some indices are based around mutation, a phenomenon unique to Welsh and other Celtic languages in which the initial consonant of a word changes systematically according to its grammatical context (e.g. *merch* → *y ferch*). These mutation features add an additional dimension of profiling that reflects a core grammatical process of Welsh.

3.4. Extraction Pipeline

The extraction pipeline processes texts in batches to compute the indices. It applies the readability analysis function to each text, extracting all defined features across linguistic classes. The pipeline is designed to handle large datasets with efficiency and resilience, making use of batch processing, retry logic (particularly to prevent UDPipe throttling), and periodic checkpointing to ensure robustness against errors and interruptions. The resulting feature vectors are stored in a JSON format for each text and can later be aggregated across text types to compile meaningful summary statistics.

[3]The full list of indices is included in the Appendix 8

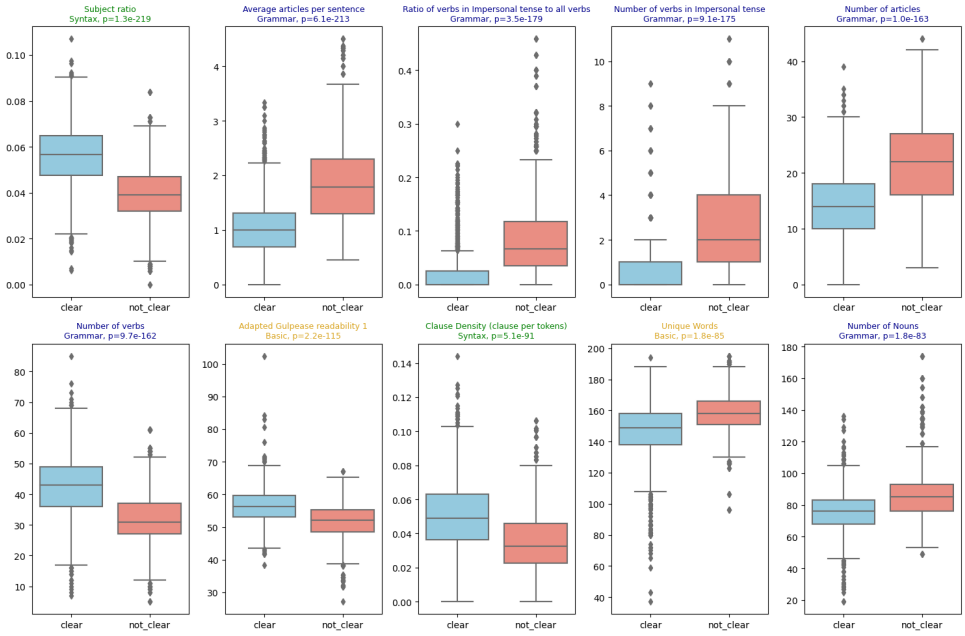

Figure 1: Top 10 significant features when profiling Cymraeg Clîr

4. Profiling Text Readability

In this section, we assess the utility of the **Proffiliadur** toolkit in analysing text complexity and readability. We evaluate the toolkit using two complementary perspectives: (i) text clarity, based on the *Cymraeg Clîr* principles, and (ii) graded language proficiency levels following the Common European Framework of Reference for Languages (CEFR).

4.1. Corpora

Cymraeg Clîr. For the clarity-based profiling experiments, we used the Welsh Text Summarisation Dataset introduced by Ezeani et al. (2022). This dataset comprises Welsh Wikipedia articles and human-generated summaries produced by native speakers. Each article–summary pair was designed so that the summary adhered to *Cymraeg Clîr* guidelines, and produce concise and readable content. The dataset contains 2,600 article-summary pairs covering a broad range of topics sourced from Welsh Wikipedia. Full articles were labelled as "not clear", while human summaries were labelled as "clear".

We acknowledge that this labelling scheme constitutes a proxy for readability rather than a direct, independently annotated measure reliant on human judgement. Traditional readability formulae such as Flesch-Kincaid estimate difficulty numerically along a graded scale and are calibrated against extensive human-annotated scores. However, in the absence of similar data in Welsh, adherence to *Cymraeg Clîr*'s stylistic principles based on the summariser's judgement is used as a proxy for readability grounded in text accessibility. While this is not exactly measuring intrinsic textual complexity, it captures similar features and provides a reasonable alternative for approximating readability.

CEFR-Level Readability. For the analysis of proficiency-related readability, we used the Welsh subset of the UniversalCEFR dataset (Imperial et al., 2025), which contains texts aligned with CEFR levels A1 to A2, as well as the dataset in (Waqar et al., 2026) containing texts with CEFR level B1 to B2, all obtained from reference textbooks published by Learn Welsh.[4] Due to the current availability of CEFR resources, C1-C2 levels were not available, thus limiting our analysis to the A and B levels (see Limitations).

4.2. Cymraeg Clîr Analysis

Proffiliadur was applied to all texts along "clear" and "not clear" labels to extract all types of indices. Then, features showing statistically significant differences between the two text types (as determined by t-tests) were examined. This analysis allowed to assess whether "not clear" texts exhibit greater structural, syntactical and lexical complexity as compared to "clear" texts. Figure 1 shows the top 10 features by statistical significance.

Analysis of the most significant features reveals linguistic contrasts between "clear" and "not clear" texts. "Not clear" texts exhibit a lower **subject ratio**, reflecting Welsh's pro-drop tendency. While inflection in Welsh grammar allows for subjects to be

[4] https://learnwelsh.cymru/learning/resource-library/?k=Coursebooks&opt=Tags

ommitted, this can obscure meaning. For example, *rheda i* (I am running) may drop the pronoun (*i*) to become *rhedaf* without subject ambiguity, whereas *rhediff hi* and *rhediff e* (she/he is running) both reduce to *rhediff*, concealing the subject. Similarly, "not clear" texts often feature **short-form verbs** including use of the **impersonal mood**, which rely on conjugations that are more typical of a literary register and might not be accessible to all readers. Although these forms provide more concise language, they go against *Cymraeg Clîr* recommendations, which favour explicit subjects and avoidance of the impersonal mood.

In contrast, "clear" texts are characterised by higher sentence counts and clause density, distributing information across several shorter sentences rather than compressing it into denser structures. This pattern also directly aligns with *Cymraeg Clîr* guidance to keep sentences short, focused on a single idea, and punctuated in ways that reflect natural pauses in speech. "Clear" texts also contain a higher overall number of verbs, likely due to a preference for **periphrastic constructions** that combine auxiliary and main verbs. Such constructions increase verbal count while promoting explicit phrasing that mirror the rhythms of spoken Welsh, promoting a more direct, conversational style rather than the compressed, literary style. This reflects *Cymraeg Clîr* principles encouraging language that feels natural to speakers: active, direct, and rooted in everyday usage.

Features such as **number of articles** and **number of punctuations** are higher in "not clear" texts, suggesting a tendency toward more complex noun phrases and clause boundaries, which are both linked to greater syntactic complexity. Finally, another *Cymraeg Clîr* guideline advises rewriting in a way that avoids mutations, especially on posters, forms, or when foreign words occur. This pattern is reflected in the data, where "not clear" texts have a statistically significant higher percentage of overall mutations, and especially soft mutations.

Finally, across the top thirty most significant features, grammatical indices dominate (53%). Surface-level and syntactic indices are 23% and 20%, respectively, with only one index in the top 30 being lexical. This distribution suggests that text clearness in Welsh depends more on grammatical and syntactic factors, especially explicit subjects, direct verb forms, and clear clause boundaries, rather than on vocabulary choice or lexical diversity.

4.3. CEFR Levels Analysis

Analysis of statistically significant features across CEFR levels (A1-B2) was conducted using a one-way analysis of variance (ANOVA) per feature, followed by Tukey Honest Significant Difference (HSD) post-hoc tests to identify statistically significant differences within level pairs.

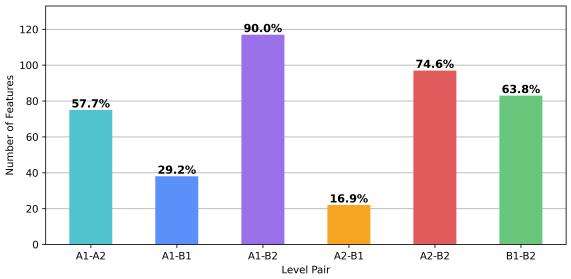

Figure 2: Significant Features per Level Pair (Tukey HSD, $p < 0.05$)

Most features were found to be statistically significant, with 130/141 (92.2%) showing differences across at least one level pair. When grouped by class, 100% of Surface-level features, 90.9% of Vocabulary features, 96.1% of Grammatical features, and 75% of Syntactic features exhibited statistically significant variation across proficiency levels.

Post-hoc pairwise comparisons revealed systematic differences in the distribution of significant features across level transitions. Figure 2 shows the percentage of significant features per level pair. The most distant pairs (A1-B2) exhibited the greatest number of discriminating features, while adjacent pairs displayed fewer. Notably, the A2-B1 pair returned the lowest number across all pairs. This is consistent with the gradual nature of proficiency progression in the CEFR framework, where adjacent levels represent incremental rather than sharply distinct development, As a result, linguistic features tend to be similar in neighbouring levels, making them harder to discriminate statistically.

Several features display a predictable monotonic increase across levels. Average Sentence Length, Ratio of verbs in Present tense (which also reflects the use of periphrastic structures for other tenses), Average Adverbs per sentence, Average Object Depth, and Word Count, all increase progressively from A1 to B2 with a statistically significant p-value. These trends align with established expectations: higher-proficiency texts tend to be longer, syntactically deeper, and lexically richer.

Not all features follow this pattern. Notably, Average soft mutations per sentence is highest in A1. While the absolute number of soft mutations is actually lowest in A1, sentence length at this level is typically very short, resulting in a relatively high per-sentence ratio. This pattern also reflects the pedagogical nature of the texts: the A1 module explicitly introduces soft mutation, leading to frequent but highly-concentrated usage, and consequently an inflated per-sentence average despite the lower overall token frequency.

Altogether, these findings suggest that the fea-

ture set is effective in distinguishing CEFR levels, capturing proficiency progression most reliably at the extremes of the scale, with diminishing discriminative power across intermediate transitions. This pattern likely reflects both the cumulative nature of linguistic complexity, as well as the design of the CEFR framework, where early and advanced stages exhibit clearer structural differences than the more gradual transitions observed between intermediate levels.

5. Cymraeg Clîr Classification

To evaluate the potential of **Proffiliadur** for automated readability assessment, we conducted a classification experiment to distinguish between texts labelled as "clear" and "not clear". Four models were employed: (i) a feature-based logistic model using **Profilliadur**'s indices, (ii) a logistic classifier trained on static FastText CBOW embeddings, (iii) a fine-tuned BritLLM model,[5] and (iv) a few-shot prompted Claude-Sonnet-4.5 model. All models are evaluated using the same five-fold stratified cross-validation splits to ensure balanced class representation and allow direct comparison. Performance across all models is reported using accuracy, precision, recall, and F1-score.

5.1. Models

5.1.1. Logistic Regression + Proffiliadur

The first model relies on a set of linguistic features extracted from each text using the toolkit indices. The goal of this model is to assess to what extent traditional linguistic features can distinguish between texts labelled as "clear" and "not clear".

Each text is represented as a fixed-length feature vector and used to train a logistic regression classifier. This model was chosen for its simplicity and interpretability, since the model weights provide a direct mapping between specific linguistic features and the predicted clearness label, allowing identification of which aspects contribute most to readability classification.

This feature-based approach serves as a lower-complexity but linguistically rich baseline against which the embedding-based and large language model (LLM) classifiers can be compared to assess the added value of semantic and contextual information, and whether the trade-off between interpretability and accuracy is worth it.

5.1.2. Logistic Regression + FastText

The second model presents the most basic embedding-based baseline using static FastText CBOW embeddings. Each text is represented as a fixed-length vector computed from the average of its word embeddings, capturing general distributional semantic information, and a logistic regression classifier is trained on these vectors.

5.1.3. Fine-tuned BritLLM

Four our third model, we fine-tuned BritLLM[6] (a pre-trained LLM) for sequence classification using Low-Rank Adaptation (LoRA) to efficiently adjust model parameters. The goal of this model is to leverage contextual embeddings and the deep representations of the transformer architecture for the task of classifying "clear" and "not clear" texts. The deep contextual representations of this model are suitable for capturing subtler patterns than those accessible to traditional-feature based methods, though losing out on interpretability.

The base BritLLM model chosen (britllm/britllm-3b-v0.1) has 3 billion parameters, but LoRA injects trainable adapters into the attention projections, with hyperparameters $r = 8$, $\alpha = 16$, $dropout = 0.05$. Training was done using batch size 4, learning rate of 1e-4 and 3 epochs, with sequences truncated at 256 tokens.

5.1.4. Few-shot prompted Claude-Sonnet-4.5

Finally, the fourth model explores few-shot prompting with Claude-Sonnet-4.5 via the OpenRouter API. Each text is evaluated within a carefully designed prompt that provides specific instructions and two examples, one for each label, illustrating the clearness criteria. The prompts guide the models to apply the *Cymraeg Clîr* principles, including guidelines like "clear texts use short, direct sentences, active verbs, simple vocabulary", and "not clear texts include technical jargon, impersonal constructions, long sentences, and abstract language". The full prompt is included in the appendix 8.

5.2. Results

As shown in Table 1, the first logistic classifier trained on **Proffiliadur**'s indices achieves strong performance across folds. These results indicate that linguistic features alone can reliably capture patterns of clarity. Examination of the learned coefficients shows that "not clear" texts contain more nouns, grammatical auxiliaries (articles, particles,

[5] https://huggingface.co/britllm/britllm-3b-v0.1

[6] Available at https://huggingface.co/britllm and https://llm.org.uk/. BritLLM is an open large language model developed by University College London, pre-trained on data covering the languages of the British Isles, including Welsh, Irish, Scottish Gaelic, and Scots.

Model	Accuracy	F1	Precision	Recall
Linguistic Features	0.943	0.944	0.930	0.959
Static Embeddings	0.841	0.844	0.827	0.861
BritLLM finetuned	**0.966**	**0.966**	**0.967**	0.966
Claude prompting	0.543	0.685	0.522	**0.994**

Table 1: Performance Metrics across Models. Best scores in bold.

prepositions), punctuations, and clauses, suggesting more diffuse information and greater lexical diversity (higher type-token ratio). In contrast, "clear" texts are characterised by more verbs, compact clause structures, broader dependency trees, and shorter overall length.

The logistic model using FastText CBOW embeddings achieves moderate performance. These results indicate that even a simple baseline using static CBOW embeddings captures substantial information about the clearness of the text, though without the interpretability of explicit linguistic features. Serving as a minimal embedding-based baseline, it provides a reference point for evaluating more sophisticated contextual and transformer-based approaches.

The fine-tuned BritLLM model benefits from deep contextual embeddings, allowing it to capture subtler semantic and syntactic patterns than the previous approaches. It achieves the highest performance, indicating a reliable distinction between labels. These results demonstrate that transformer-based models can be fine-tuned to perform effectively in Welsh readability classification.

The prompted Claude model exhibited very high recall for "not clear" texts, correctly identifying nearly all examples, but it exhibited substantial overprediction of this label, resulting in a low precision of 0.522 and accuracy of 0.543. These results suggest that the model did not generalize clearness rules based on the prompt and few-shot examples, defaulting extensively to "not clear". The model overgeneralised surface complexity and formal register cues, leading to overprediction of "not clear". This behaviour is likely to stem from asymmetries in the prompt criteria, in which features of "not clear" were interpreted as hard limits, whereas features of "clear" were interpreted as soft tendencies, resulting in a bias towards "not clear". Additionally, limited few-shot examples, limited knowledge of Welsh of the model, or imperfect internalisation of *Cymraeg Clîr* principles may have contributed to the limited performance.

While few-shot prompting provides a baseline using a large model without fine-tuning, the findings indicate that prompting alone may be insufficient for readability tasks requiring in-depth judgement of semantic and syntactic features in under-represented languages. Future experiments could explore modifications to the prompting structure such as changing the prompt language, providing additional examples in the few-shot setting, or using stepwise reasoning to improve performance.

6. Discussion

The analyses presented in this paper demonstrate both the potential and the challenges in assessing text clarity or readability in Welsh. The dataset profiling in Sec. 4 shows that structural and syntactic characteristics are strong indicators of readability, consistent with guidelines from *Cymraeg Clîr* that emphasize clarity in morphological constructions and spoken-style syntax. In particular, features such as explicit subjects, avoidance of the impersonal mood, the use of periphrastic rather than concise verb forms, and shorter clauses strongly correlate with clearer writing. These results suggest that readability in Welsh is shaped less by lexical simplicity and more by morphosyntactic structure, especially how far a text departs from the condensed and formal structures more typical of literary Welsh.

The prominence of grammatical and syntactic features amongst significant indices demonstrate the value and utility of the **Proffiliadur** toolkit in providing interpretable metrics across multiple linguistic dimensions, and in capturing patterns that may be overlooked by generic readability metrics. It provides quantifiable evidence for long-standing stylistic intuitions in *Cymraeg Clîr*: that clear Welsh should resemble spoken usage more than written tradition, favouring shorter sentences, explicit clause boundaries, and natural syntactic rhythm reflected in punctuation. Importantly, these findings reinforce that Welsh readability is not merely a matter of word choice or vocabulary, but rather how grammatical constructions shape comprehension.

The classification experiments in Sec. 5 further validate these insights. The feature-based logistic regression model, which relies exclusively on interpretable linguistic metrics from the toolkit, achieves strong performance (average $F_1 = 0.94$), nearly equivalent to the fine-tuned BritLLM (average $F_1 = 0.966$), while retaining full interpretability with minimal computational resources. This suggests that much of what enables LLMs to distinguish "clear" from "not clear" texts can be captured through traditional linguistic features, which is a key result for low-resource languages like Welsh where large-scale fine-tuning is often impractical due to the lack of data resources.

The simple CBOW embedding baseline performs modestly (average $F_1 = 0.844$), indicating that distributional semantics capture useful information, but with lower precision and limited interpretability. In contrast, few-shot prompting Claude-Sonnet-4.5 underperformed expectations, scoring substan-

tially lower than even the CBOW baseline (average $F_1 = 0.685$) and demonstrating strong bias toward "not-clear" predictions. This suggests that few-shot prompting, while convenient, was insufficient for reliably applying *Cymraeg Clîr* principles.

These results highlight a clear trade-off between interpretability and model complexity. While fine-tuned BritLLM achieves the highest predictive accuracy, the marginal gain over the feature-based model is moderate (~3% difference in F1 and accuracy), whereas the feature-based model offers fully interpretable outputs and lower computation cost. From a practical perspective, this reinforces the value of the **Proffiliadur** toolkit in providing robust, linguistically grounded measures of readability in Welsh, with potential to diagnose potential barriers to comprehension, and support evidence-based decisions for text simplification.

7. Conclusion and Future Work

This paper introduced *Proffiliadur*, a Welsh language text profiling toolkit that computes 141 linguistic indices across surface, lexical, morphological, and syntactic levels. The toolkit addresses specific challenges of Welsh, such as mutation handling, rich morphology, and syntactic variation, providing interpretable features for text analysis.

Profiling of texts following *Cymraeg Clîr* principles and across two CEFR proficiency levels demonstrated that grammatical and syntactic features, rather than purely lexical ones, are most predictive of text readability in Welsh. Feature-based classification models built with these indices achieved high performance ($F_1 = 0.94$), nearly matching a fine-tuned large language model ($F_1 = 0.97$). This finding suggests that carefully designed linguistic features can rival complex neural architectures for specific languages tasks, especially in low-resource settings where interpretability, accountability, access to data, and cost-efficiency are highly important.

Beyond technical performance, the results offer empirical evidence of the stylistic and linguistic patterns associated with *Cymraeg Clîr* principles. Specifically, features such as explicit subjects, avoidance of the impersonal mood, and preference for shorter clause constructions, are both statistically and computationally linked with clear writing.

Future work will extend **Proffiliadur** with discourse-and semantic-level indices, and evaluate it on a broader range of public-sector, educational, medical and technical texts. The toolkit and models will be made openly available to promote reproducible research and support the development of accessible Welsh language technologies.

Limitations

Proffiliadur currently targets standard written Welsh and does not yet handle dialectal or colloquial varieties. As the system depends on CyTag2, any tokenisation or tagging errors propagate downstream to feature counts, and named-entity features remain limited by its imperfect proper-noun detection. For instance, if the POS-tagging system misidentifies 'ei' as a conjugation of 'mynd' (go) as opposed to the possessive pronoun due to errors in the disambiguating process, this would affect downstream feature extraction. The toolkit focuses on sentence-level measures and lacks discourse-level metrics, and syllable-based indices cannot yet be computed due to the absence of a Welsh syllable splitter. The low-resource context also prevents inclusion of popular readability formulas such as Flesch–Kincaid or Gunning Fog, which are language specific and require syllable counts.

Although both evaluation datasets are Welsh-specific, corpus diversity is still limited: the Summarisation dataset's clear/not-clear mapping may not perfectly represent readability contrasts, and the CEFR corpus includes only A1-B2, since other levels are not available yet. Validation against human judgement such as user studies have not been conducted yet to confirm alignment with reader perception. Future work will expand corpus coverage, develop syllable-based and discourse metrics, and improve robustness across domains and text varieties.

8. Bibliographical References

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: a multilingual analyzer on multiple levels of language for readability assessment.

Daniel Cunliffe, Andreas Vlachidis, Daniel Williams, and Douglas Tudhope. 2022. Natural language processing for under-resourced languages: Developing a welsh natural language toolkit. *Computer Speech & Language*, 72:101311.

Mrinmoy Dalal et al. 2023. Language agnostic readability assessments. In *Proceedings of the 2023 IEEE Conference on Software Engineering and Computer Engineering (CSCE)*, pages 252–257. IEEE.

Ignatius Ezeani, Mahmoud El-Haj, Jonathan Morris, and Dawn Knight. 2022. Introducing the Welsh text summarisation dataset and baseline systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5097–5106, Marseille, France. European Language Resources Association.

D. B. Friedman, L. Hoffman-Goetz, and J. F. Arocha. 2006. Health literacy and the world wide web: Comparing the readability of leading incident cancers on the internet. *Medical Informatics and the Internet in Medicine*, 31(1):67–87.

Itziar Gonzalez Dios. 2016. *Readability Assessment and Automatic Text Simplification: The Analysis of Basque Complex Structures*. Ph.D. thesis, University of the Basque Country, Basque Language and Communication Department. CC BY-NC-SA 4.0.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. ArXiv preprint.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Dawn Knight, Steven Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, Enlli Môn Thomas, Alex Lovell, Jonathan Morris, Jeremy Evas, Mark Stonelake, Laura Arman, Joshua Davies, Ignatius Ezeani, Steven Neale, Jennifer Needs, Scott Piao, Mair Rees, Gareth Watkins, Lowri Williams, Vignesh Muralidaran, Bethan Tovey-Walsh, Laurence Anthony, Tom Cobb, Margaret Deuchar, Kevin Donnelly, Michael McCarthy, and Kevin Scannell. 2020a. Corcencc: Corpws cenedlaethol cymraeg cyfoes – the national corpus of contemporary welsh. *Cardiff University*.

Dawn Knight, Steven Morris, Tony Fitzpatrick, Paul Rayson, Irena Spasić, and Enlli M. Thomas. 2020b. The national corpus of contemporary welsh: Project report | y corpws cenedlaethol cymraeg cyfoes: Adroddiad y prosiect. ArXiv preprint arXiv:2010.05542.

Dawn Knight, Stuart Morris, Bethan Tovey-Walsh, Tom Fitzpatrick, and Laura Anthony. 2020c. Yr amliadur: Frequency lists for contemporary welsh. Technical report, Cardiff University.

Ivan Kraus. 2025. Predicting readability of czech legal writing using linguistic features. Bachelor's thesis defended on June 9, 2025.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2024. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Tim Loughran and Bill McDonald. 2015. Measuring readability in financial disclosures. *Journal of Financial Economics*, 118(2):258–278.

P. Lucisano and Maria E. Piemontese. 1988. Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città*, 39(3):110–124. In Italian.

Emilio Matricciani. 2023. Readability indices do not say it all on a text readability. *Analytics*, 2(2):296–314.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Steven Neale, Kevin Donnelly, Gareth Watkins, and Dawn Knight. 2018. Leveraging lexical resources and constraint grammar for rule-based part-of-speech tagging in welsh. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3946–3954, Miyazaki, Japan. European Language Resources Association (ELRA).

OECD. 2011. National survey of adult skills 2011. https://www.oecd.org/skills/nationalsurveyofadultskills.htm. Accessed 13 October 2025.

Lloyd Lois Antonie Reyes, Michael Antonio Ibañez, Ranz Sapinit, Mohammed Hussien, and Joseph Marvin Imperial. 2022. A baseline readability model for cebuano. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. Association for Computational Linguistics. Accepted to BEA Workshop at NAACL 2022.

Tanya Serry, Tonya Stebbins, Andrew Martchenko, Natalie Araujo, and Brigid McCarthy. 2022. Improving access to covid-19 information by ensuring the readability of government websites. *Health Promotion Journal of Australia*, 34(1):68–72.

Milan Straka and Jana Straková. 2017. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphosyntactic tagging, lemmatization and dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Cíntia Matsuda Toledo, Sandra Maria Aluísio, Leandro Borges dos Santos, Sonia Maria Dozzi Brucki, Eduardo Sturzeneker Trés, Maira Okada de Oliveira, and Letícia Lessa Mansur. 2018. Analysis of macrolinguistic aspects of narratives from individuals with alzheimer's disease, mild cognitive impairment, and no cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:31–40.

S. Uchida. 2025. Assigning cefr-j levels to english learners' writing. *Journal of Second Language Writing*, S2772766125000205.

Tiffany M. Walsh and Teresa A. Volsko. 2008. Readability assessment of internet-based consumer health information. *Respiratory Care*, 53(10):1310–1315.

Zhi Wang, Yifan Zhang, Zhi Li, and Zhiwei Liu. 2019. Readability assessment of textbooks in low resource languages. volume 61, pages 23109–23121.

Eeshan Waqar, Jonathan Davies, Dawn Knight, and Fernando Alva-Manchego. 2026. Cefr-cymraeg: A dataset and baseline models for language proficiency assessment in welsh. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma, Mallorca. European Language Resources Association.

Welsh Government. 2017. Cymraeg 2050: A million welsh speakers. https://www.gov.wales/cymraeg-2050-welsh-language-strategy. Accessed 13 October 2025.

Welsh Government. 2018. Welsh language technology action plan. https://www.gov.wales/welsh-language-technology-action-plan. Accessed 13 October 2025.

C. Williams. 1999. *Cymraeg Clir*. Cyngor Gwynedd.

Full list of Features

Table 2: Full List of Toolkit Indices, with Subclasses and explanations

Feature	Subclass	Description / Notes
General Text Statistics and Readability Indices		
Sentence Count	Basic	Total number of sentences in the text.
Token Count	Basic	Total number of tokens.
Number Count	Basic	Number of numeric tokens.
Word Count	Basic	Count of words excluding punctuation.
Average Sentence Length	Basic	Average number of tokens per sentence.
Unique Words	Basic	Number of unique word types.
TTR (Type Token Ratio)	Basic	Ratio of unique types to total tokens.
TTR (Token Type Ratio)	Basic	Alternative form of TTR.
Adapted Gulpease readability (1–3)	Readability	Readability scores.
Lexical Profiling (a1/a2/b1/K1–K6/stopwords)		
a1 - unique matches	Lexical	Count of unique words in the A1 WJEC/CEFR band.
a1 - total appearances	Lexical	Total occurrences of A1 words.
a2 - unique matches	Lexical	Unique words in A2 WJEC/CEFR band.
a2 - total appearances	Lexical	Total appearances in A2 band.
b1 - unique matches	Lexical	Unique words in B1 WJEC/CEFR band.
b1 - total appearances	Lexical	Total appearances in B1 band.
K1–K6+ (unique matches)	Lexical	Unique word counts by frequency level.
K1–K6+ (total appearances)	Lexical	Total appearances by frequency level.
K1–K6+ (ratio)	Lexical	Proportion of words in each band.
cy_ataleiriau – unique stopwords	Lexical	Number of distinct Welsh stopwords.
cy_ataleiriau – total appearances	Lexical	Total stopword occurrences.
cy_ataleiriau_stopwords – unique stopwords	Lexical	Distinct stopwords in extended list.
cy_ataleiriau_stopwords – total appearances	Lexical	Total count for extended stopwords.
cy_ataleiriau – ratio to all words	Lexical	Stopword proportion.
cy_ataleiriau_stopwords – ratio to all words	Lexical	Ratio for extended stopword set.
Part-of-Speech Statistics		
Coverage of POS tags	Grammar	Percentage of possible POS tags represented.
Number of types of POS Tags	Grammar	Count of unique POS tags present.
Lemma Count	Grammar	Number of lemmatized tokens.
Number of verbs	Grammar	Total verb count.
Unique verb lemmas	Grammar	Unique verb lemmas.
Average verb count per sentence	Grammar	Mean verbs per sentence.
Number / Ratio of verbs by tense (Present, Past, Imperfect, etc.)	Grammar	Distribution of verb tenses.
Number of verb nouns	Grammar	Count of verb-noun forms.
Number of Nouns	Grammar	Total noun count.
Average nouns per sentence	Grammar	Mean nouns per sentence.
Unique noun lemmas	Grammar	Unique noun lemmas.
Gender (Masculine/Feminine)	Grammar	Count and ratio of nouns by gender.
Number (Singular/Plural)	Grammar	Count and ratio of nouns by number.
Proper Nouns	Grammar	Count and ratio of proper nouns.

Feature	Subclass	Description / Notes
Number / Average adjectives, adverbs, prepositions, numerals, articles, conjunctions, pronouns, determiners, particles	Grammar	Counts and ratios of key grammatical categories.
Coordinating vs Subordinating conjunctions	Grammar	Counts and ratios per subtype.
Mutation Features (Welsh-specific)		
Number / Average mutations per sentence	Grammar	Total and average counts of mutations.
Soft / Nasal / Aspirate / h-prothesis mutations	Grammar	Subtypes of mutation with counts and averages.
Syntax - Dependency Tree Metrics		
Max / Average Tree Depth	Syntax	Measures of syntactic complexity.
Max / Average Tree Breadth	Syntax	Structural breadth of dependency trees.
Max branching factor	Syntax	Degree of syntactic branching.
Average head distance	Syntax	Mean dependency head distance.
Subject ratio / Object ratio	Syntax	Proportion of explicit subjects/objects to total tokens.
Average subject/object/verb depth	Syntax	Mean syntactic depth of core arguments.
Left-right branching ratio	Syntax	Directionality in dependency trees.
Clause count / density	Syntax	Total and relative clause counts.
Coordinated conjuncts	Syntax	Count of conjunctive phrases.
Auxiliaries	Syntax	Auxiliary verb counts.
Average leaf ratio	Syntax	Average ratio of leaves in parse trees.

Full Prompt for Claude

```
You are an expert in evaluating the clarity of Welsh writing using the
    Cymraeg Clir (Clear Welsh) principles.

  Your task: Decide whether the following Welsh text is 'clear' or 'not
    clear' for an average reader, based on the guidelines below.

  Clear (``clir''):
  - Short, direct sentences (around 20 words or fewer)
  - Use everyday Welsh words, not overly formal or archaic language
  - Include personal pronouns (fi, chi, ni) and simple contractions (e.g.,
    ``dyw'r'' not ``yw yr'')
  - One main idea per paragraph
  - Active verbs and plain sentence structure
  - A text can still be ``clear'' even if it has some long sentences or
    technical words, as long as the main ideas are easy to understand for an
    average reader.


  Not clear (``aneglur''):
  - Long, complex, or passive sentences
  - Use archaic, impersonal, or bureaucratic language
  - Overly abstract or technical vocabulary
  - Mixed tone or unnecessarily complicated phrasing

  Here are two examples:

  Example A (clear):
  Edrychwch o'ch cwmpas. Welwch chi rywbeth i'w ddarllen? Oes print i'w weld?
    Mae print ym mhobman y dyddiau hyn ac rydym yn gorfod ei ddarllen a'i
    ddeall yn sydyn. Mae llawer ohono'n llawn jargon, geiriau technegol,
    geiriau hen ffasiwn a brawddegau hir a chymhleth. Cofiwch chi, mae'r rhai
    sy'n sgrifennu hysbysebion wedi deall y gyfrinach. Brawddegau byr sy'n
    dal eich sylw sydd ganddyn nhw ran amlaf; neges sy'n hawdd i'w deall.
  Label: clear

  Example B (not clear):
  Yn y byd cyfoes, amgylchynir ni gan ddeunyddiau argraffedig y disgwylir i'r
    cyhoedd eu darllen a'u dehongli gyda brys anarferol er eu bod, yn aml, yn
    llawn ffiloreg a geirfa dechnegol a thraddodiadol yn ogystal
    chystrawennau anhylaw sy'n cyflwyno ystyron astrus. Ar y llaw arall,
    gellid dadlau bod y dylunwyr a'r crewyr hynny sy'n gyfrifol am
    hysbysebion wedi datrys y gyfrinach trwy arfer brawddegau byrion i gyfleu
    cenadwri y mae dealltwriaeth ohoni yn haws.
  Label: not clear

  Now assess the following text:

  {text}

  Reply with one label only:
  - clear
  - not clear
```

Figure 3: Full prompt used for Claude.