LLMs Know More Than Words: A Genre Study with Syntax, Metaphor & Phonetics

Anonymous ACL submission

Abstract

Large language models (LLMs) demonstrate remarkable potential across diverse language-related tasks, yet whether they capture deeper linguistic properties—such as syntactic structure, phonetic cues, and metrical patterns—from raw text remains unclear. To analysis whether LLMs can learn these features effectively and apply them to important nature language related tasks, we introduce a novel multilingual genre classification dataset derived from Project Gutenberg, a large-scale digital library offering free access to thousands of public domain literary works, comprising thousands of sentences per binary task (poetry vs. novel; drama vs. poetry; drama vs. novel) in six languages (English, French, German, Italian, Spanish, and Portuguese). We augment each with three explicit linguistic feature sets (syntactic tree structures, metaphor counts, and phonetic metrics) to evaluate their impact on classification performance. Experiments demonstrate that although LLM classifiers can learn latent linguistic structures either from raw text or from explicitly provided features, different features contribute unevenly across tasks, which underscores the importance of incorporating more complex linguistic signals during model training.

1 Introduction

Large Language Models are extensively investigated in text generation, translation, summarization, and many kinds of language-related tasks (Chkirbene et al., 2024; Li et al., 2024), along with widespread adoption across the social sciences and liberal arts (Thapa et al., 2025; Ziems et al., 2024; Wang et al., 2024). Meanwhile, researchers have begun exploring their intersections with linguistic science (Muñoz-Ortiz et al., 2024; Rosenfeld and Lazebnik, 2024). Some studies have examined whether synthetic texts generated by LLMs differ fundamentally from human-authored texts (Muñoz-Ortiz et al., 2024), while others have sought to uncover novel linguistic patterns emerging in AI outputs (Kuwanto et al., 2024), which can be used both for developing robust methods to detect LLM-generated content (Park et al., 2025) and for advancing automated language generation. However, despite the crucial role that linguistic features play in natural language, relatively little research has been conducted in this area. The ability of models to comprehend and leverage these features for downstream tasks has long been overlooked, leaving a significant gap in our understanding of how linguistic structures can enhance model performance. Therefore, a dataset that combines sufficient scale, linguistic complexity, and multilingual coverage is essential for training and evaluating models on these fundamental tasks. Moreover, a rigorous and scientific framework is also needed to assess how well models understand latent linguistic features.

We argue that genre classification offers the most effective framework for evaluating a model's grasp of linguistic features. Literary scholars have long recognized that genres are defined by distinctive linguistic patterns: they vary in syntactic tree structures (Dell'Oglio et al., 2018; Brigadoi, 2021), in the use of metaphor, and in phonetic characteristics such as metrical patterns. By focusing on these core differences, genre classification directly tests a model's ability to capture the latent structures that distinguish one form of writing from another.

Therefore, in this paper, we construct a multilingual genre dataset comprising novels, poetry, and drama from Project Gutenberg, a large-scale digital library offering free access to thousands of public domain literary works and always used in linguistic research, to investigate how latent linguistic structures influence a model's ability to distinguish between literary forms. For each language (English, French, German, Italian, Spanish, and Portuguese) we extract thousands of sentences for

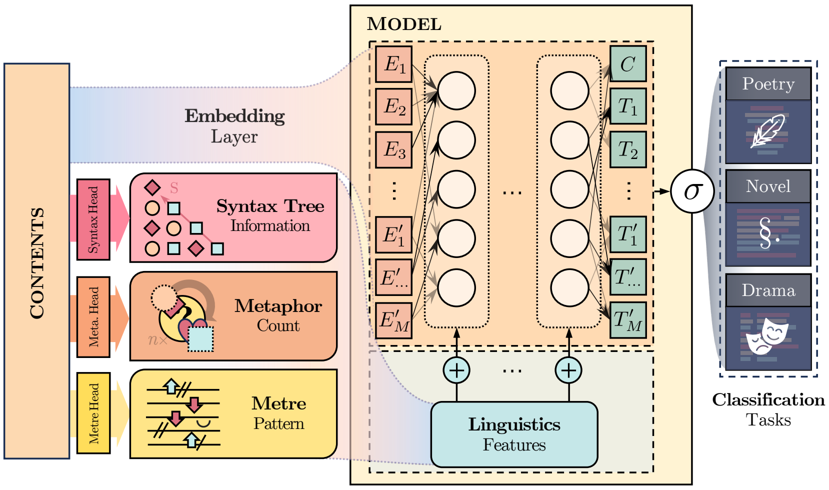


Figure 1: Overview of our method. We extract various types of linguistic information from raw text and integrate them with the original sentences. The model is then trained to embed these enriched inputs uniformly into the latent space, thereby enhancing performance on genre classification tasks.

three binary classification tasks (poetry vs. novel, drama vs. poetry, and drama vs. novel). We then fine-tune models using a naive approach as well as with three explicit linguistic features: syntactic tree depth, metaphor frequency, and phonetic regularity. Results show consistent improvements across most metrics when incorporating linguistic features. Experiments demonstrate that incorporating phonetic features into the training process yields consistent performance improvements across most models. The result indicate that the classifiers effectively learn latent linguistic features in the sentences and that different feature types contribute variably depending on the specific classification task.

In conclusion, this paper makes three key contributions. **First**, we introduce a novel multilingual literary genre classification dataset derived from Project Gutenberg, a large-scale digital library offering free access to thousands of public domain literary works, spanning six languages and three binary genre distinctions, which serves as a fundamental tool for evaluating LLMs' sensitivity to latent linguistic features. **Second**, we systematically augment this dataset with three types of linguistically motivated features, providing a framework to assess how explicit linguistic cues affect model performance. **Third**, through extensive experiments, we demonstrate that LLMs can both infer and benefit from these latent structures, showing varied but consistent performance gains, which underscores the importance of linguistically enriched signals in advancing LLM-based literary and language understanding.

Based on the experimental results, we derive the following insights:

(1) **Incorporating latent linguistic structure information into model training can enhance performance on various language tasks that involve complex linguistic features.** This approach offers a promising pathway to align model-generated content more closely with human language, potentially bridging the subtle gaps that often arise between AI-generated and human-authored texts.

(2) **Different linguistic features contribute unevenly to model learning.** Grammatical cues such as syntactic tree structures are relatively easy for models to learn, whereas phonetic patterns are more subtle and difficult to capture, as they are closely tied to sound and thus inherently multimodal. Interestingly, these more latent features tend to yield greater performance improvements, suggesting that supplying models with deeper, multimodal linguistic signals may significantly boost their capacity to handle complex language tasks.

2 Related Work

In this section, we review related work on the interaction between linguistic features and large language models (LLMs), and discuss how the task of genre classification fits into this broader research landscape.

Linguistic Features and LLMs. Interdisciplinary research at the intersection of large language models (LLMs) and linguistics has only begun to emerge in recent years. Early studies focused on evaluating whether encoder-only models, such as BERT (Koroteev, 2021), possess the capacity to understand the construction grammar of sentences (Tayyar Madabushi et al., 2020). Subsequent work has involved the use of LLMs to annotate sentence and paragraph structures (Yu et al., 2024; Ljubešić and Kuzman, 2024), as well as comparing the linguistic characteristics of sentences generated by decoder-only models, such as LLaMA (Touvron et al., 2023), Falcon (Harabagiu et al., 2000) and Mistral (Jung et al., 2010)—with those written by humans (Muñoz-Ortiz et al., 2024). Researchers have also begun to investigate whether LLMs are capable of understanding phonetic (Ballier et al., 2023; Doshi et al.) and morphological information (Anderson et al., 2025; Asgari et al., 2025) embedded in sentences. However, no study has exclusively focused on how linguistic characteristics can be measured through genre classification tasks. Moreover, existing datasets and benchmarks have primarily concentrated on English texts and often lack rigorous validation based on linguistic theory. These gaps highlight the need for a multilingual and theoretically grounded evaluation pipeline to assess how effectively models classify genres and to provide deeper insights into the role of underlying linguistic structures in this fundamental task.

Genre Classification. Genre classification, by contrast, has long been explored through computational methods. Early efforts focused on using statistical learning techniques to automatically classify genres based on the lexical features of documents (Lee and Myaeng, 2002). In the era of large language models (LLMs), most research has centered on constructing datasets from a functional perspective (Egbert et al., 2015; Kuzman et al., 2022; Sharoff, 2018; Kuzman et al., 2023), distinguishing between genres such as news articles, speeches, fiction, and song lyrics. Other work has examined emotional or stylistic dimensions (Hicke and Mimno, 2025). However, the prevailing methodologies still rely largely on word-level features (Bhattacharjee et al., 2024), and tend to focus on monolingual or limited multilingual contexts, often overlooking the broader multilingual dimension (Kuzman et al., 2022).

3 Method

In this section, we introduce the construction of our dataset, outline our problem formulation, and explain how linguistic features are incorporated into the training process.

3.1 Data Creation

Our dataset is derived from Project Gutenberg, a digital repository of public-domain e-books that are free from copyright restrictions. We sampled approximately 1500-3000 sentences for each of six languages, English (EN), French (FR), German (DE), Spanish (ES), Italian (IT), and Portuguese (PT), across three canonical literary genres: drama, poetry, and the novel. The exact number depends on the scale of the available raw text.This tripartite genre classification is well-established in literary scholarship. In total, the dataset comprises roughly 45,000 sentences. For each language–genre subset, we split the data into training and testing sets using an 80/20 ratio.

Next, we constructed three binary classification tasks—poetry vs. novel, poetry vs. drama, and novel vs. drama—by pairing each two genres. This setup enables us to more clearly investigate how different linguistic features influence genre classification across varying levels of structural and stylistic contrast. A detailed breakdown of the dataset distribution is provided in Table 1.

Table 1: Dataset statistics by genre and language.

Language	Drama	Poetry	Novel
EN	1625	3367	2633
FR	2313	2092	2397
DE	2528	2443	3481
ES	2423	2795	3102
IT	1912	2474	2836
PT	1658	1530	2734

3.2 Embedding Linguistic Features

As mentioned in Introduction1, we hypothesize that several linguistic features may assist LLMs in solving the genre classification task. Lyrical genres such as poetry exhibit distinctive linguistic characteristics compared to more conventional narrative forms like plays and novels. Novels, in turn, are inherently more narrative-driven than either drama or poetry, and thus possess features that further distinguish them. Our classification approach builds upon these core linguistic and structural differences. We focus on three main features to embed during training.

A *syntax tree* represents the syntactic structure of a sentence. Initially introduced by Noam Chomsky in his seminal work *Syntactic Structures* (Chomsky,

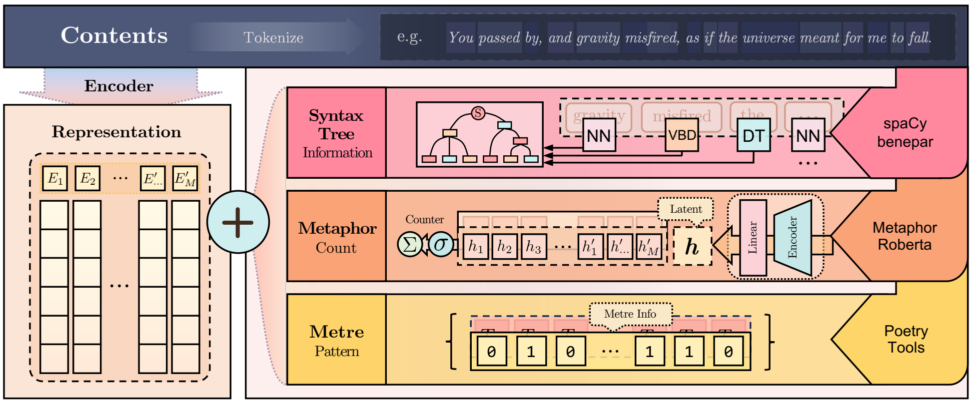


Figure 2: Detailed process of our method. We extract syntactic tree information and metrical patterns using two well-established natural language processing tools: spaCy-Benepar and PoetryTools. Metaphor counts are obtained using Metaphor RoBERTa, a state-of-the-art pretrained model designed to detect word-level metaphor usage. During training, we integrate the original sentence with the extracted linguistic features to enhance the model's performance.

2002), it has since become a foundational concept in modern linguistics (Newmeyer, 2023). We incorporate syntactic information by providing both the *syntax tree depth* and the *depth-to-length ratio* (i.e., tree depth divided by sentence length).

We examine a representative lexical feature: *metaphor count*. This is particularly useful for capturing stylistic and semantic distinctions. Metaphors reflect the density of figurative language, which tends to be more prevalent in expressive genres like poetry. We employ a metaphor detection model (Wachowiak et al., 2022) to count the number of metaphors per sentence, thereby gaining insight into the lexical choices that characterize each genre.

Phonetic features are among the most significant markers for distinguishing poetry from other literary genres (Žirmunskij, 2016). These are most commonly realized through *metre*—a system by which poets organize stress patterns, scansion, and metrical feet. The presence of metre is widely acknowledged as a key factor differentiating poetry, verse drama, and poetic prose from other genres. To evaluate whether metre is present in a sentence, we use `Poetry Tools`, a well-established Python library for analyzing metrical patterns.

Our task is defined as follows: for each binary classification instance, given an input I and an encoder-only model $E_\phi(\cdot)$, we aim to predict the label $\hat{y} \in \{0,1\}$ by computing:

$$\hat{y} = \mathbf{1}(f_\theta(E_\phi(I)) > 0.5), \tag{1}$$

where $f_\theta(\cdot)$ denotes the classification head built on top of the encoder, which may consist of a linear layer, a multi-layer perceptron, or any other parameterized transformation.

Originally, the input is defined as $I = S$, where S is the raw sentence. To investigate whether linguistic features can enhance classification performance, we construct enriched inputs by appending additional feature representations to the sentence. Specifically, we define:

$$I = S \oplus F, \tag{2}$$

where F denotes a feature vector encoding one of the linguistic features, and $\oplus$ represents concatenation in the input space. We experiment with the following features:

- **Syntax Tree Information:** represented as a tuple (d, r), where d is the syntax tree depth and $r = \frac{d}{|S|}$ is the depth-to-length ratio.
- **Metaphor Count:** a scalar feature m extracted by a pre-trained metaphor detection model, indicating the number of metaphorical tokens in S.
- **Metre Pattern:** To extract phonetic rhythm, we compute a binary stress pattern vector:

$$MP(S) = [m_1, m_2, \ldots, m_n], \quad m_i \in \{0,1\}, \tag{3}$$

where $m_i = 1$ indicates a stressed syllable and $m_i = 0$ indicates an unstressed syllable for the i-th syllable in sentence S.

Each feature is appended to the sentence embedding before being passed into the encoder. By comparing model performance on the baseline input $I = S$ and enriched versions $I = S \oplus F$, we aim to assess whether these linguistic features improve genre classification performance across different language-model pairs.

4 Experiments

In this section, we present the details of our experimental setup, report the results obtained, and provide a discussion of our findings.

4.1 Experiment Settings

Baseline We fine-tune several pre-trained BERT-based models for genre classification tasks. As baseline models, we selected BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and Metaphor-RoBERTa (Wachowiak et al., 2022). BERT and RoBERTa are widely recognized as robust encoder-only architectures and serve as standard baselines across numerous NLP tasks. DistilBERT, a lighter and faster variant of BERT, retains most of its language understanding capabilities, making it especially suitable for assessing how efficiently smaller models can learn linguistic features. Additionally, we incorporate Metaphor-RoBERTa, a model pre-trained specifically to detect metaphor usage, to investigate whether metaphor-aware pre-training enhances genre classification performance and impacts the effectiveness of our feature-enriched fine-tuning approach.

Dataset and Training We trained each model on the dataset described in Data Creation3.1, using a traditional supervised fine-tuning paradigm. For every binary genre classification task (e.g., poetry vs. novel), we input each sentence—or feature-augmented sentence—into the model, and optimize the classification loss based on the ground truth labels. We repeated this process for each of the three binary classification tasks across six languages and four models, both with and without additional linguistic features (syntax tree depth, metaphor count, and metre pattern).

Evaluation Finally, we compute the F1 scores for both genres in each binary classification task, yielding an F1 score pair $(F1_x, F1_y)$ for every combination (t, l, m), where $t \in T = \{t_1, t_2, \ldots, t_n\}$ represents the task set, $l \in L = \{l_1, l_2, \ldots, l_p\}$ denotes the language set, and $m \in M = \{m_1, m_2, \ldots, m_q\}$ is the model set. The F1 score for each class is calculated as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

where Precision and Recall are defined in the standard way based on true positives, false positives, and false negatives.

To estimate the overall effectiveness of each method across languages, we compute the macro-average of the F1 scores for each task–model pair across all languages:

$$\overline{F1}_{t,m} = \frac{1}{|L|} \sum_{l \in L} \frac{F1_x^{(t,l,m)} + F1_y^{(t,l,m)}}{2}. \tag{5}$$

This metric provides a balanced view of model performance by averaging scores for both genres and smoothing across language variations.

4.2 Results

Table 2 presents the F1 score pairs obtained for each binary classification task across six languages (EN, FR, DE, ES, IT, PT) and four baseline models (BERT, DistilBERT, RoBERTa, and Metaphor-RoBERTa). Each cell reports the F1 scores for the two genres involved in a task.

Table 3 summarizes the changes in F1 scores when incorporating three distinct linguistic features into the baseline models across the three genre classification tasks and across the languages. The detailed results are in Appendix A.

Results Across Languages. Across all languages, English and Portuguese consistently yield the highest performance across tasks and models, whereas French shows the greatest variability and relatively lower scores—especially with BERT and DistilBERT. We believe several factors contribute to this pattern:

(1) Pretraining bias: Most models are pretrained predominantly on English corpora, giving them stronger genre discrimination ability in English. They likely encounter frequent syntactic and lexical patterns in English during pretraining, making classification tasks especially easier.

(2) Linguistic proximity between genres: In French, poetry and novels exhibit syntactic and stylistic structures that are more similar than in English. As shown in Figure 3, French poetry and novel texts are less separable based on syntactic tree depth and depth-to-length ratio compared to English, making differentiation more challenging.

Table 2: F1 score statistics by genre, model and language (Percentage). We also compute the average result across all languages.

Note: P-Poetry, N-Novel, D-Drama.

Language	**EN**			**FR**			**DE**		
	P / N	P / D	N / D	P / N	P / D	N / D	P / N	P / D	N / D
BERT	0.97 / 0.97	0.97 / 0.89	0.90 / 0.67	0.50 / 0.40	0.75 / 0.71	0.76 / 0.67	0.73 / 0.78	0.72 / 0.70	0.78 / 0.75
DistilBERT	0.97 / 0.97	0.97 / 0.89	0.89 / 0.68	0.76 / 0.59	0.75 / 0.71	0.75 / 0.71	0.74 / 0.72	0.75 / 0.68	0.77 / 0.72
RoBERTa	0.94 / 0.92	0.93 / 0.73	0.90 / 0.68	0.78 / 0.82	0.76 / 0.71	0.76 / 0.71	0.60 / 0.70	0.66 / 0.63	0.78 / 0.76
Metaphor RoBERTa	0.94 / 0.92	0.92 / 0.69	0.89 / 0.67	0.77 / 0.82	0.75 / 0.75	0.79 / 0.68	0.78 / 0.82	0.79 / 0.75	0.80 / 0.79
Language	**ES**			**IT**			**PT**		
	P / N	P / D	N / D	P / N	P / D	N / D	P / N	P / D	N / D
BERT	0.77 / 0.78	0.74 / 0.70	0.77 / 0.70	0.78 / 0.81	0.81 / 0.75	0.80 / 0.69	0.80 / 0.84	0.75 / 0.64	0.75 / 0.67
DistilBERT	0.77 / 0.78	0.75 / 0.70	0.75 / 0.69	0.76 / 0.81	0.82 / 0.76	0.80 / 0.70	0.79 / 0.83	0.73 / 0.63	0.76 / 0.67
RoBERTa	0.81 / 0.83	0.71 / 0.63	0.79 / 0.69	0.83 / 0.86	0.82 / 0.79	0.82 / 0.71	0.87 / 0.89	0.75 / 0.67	0.81 / 0.73
Metaphor RoBERTa	0.80 / 0.82	0.78 / 0.77	0.82 / 0.71	0.82 / 0.85	0.80 / 0.78	0.82 / 0.72	0.86 / 0.89	0.77 / 0.68	0.81 / 0.72
Language	**Average**								
	P / N	P / D	N / D						
BERT	0.76 / 0.76	0.79 / 0.73	0.79 / 0.69						
DistilBERT	0.80 / 0.78	0.79 / 0.73	0.79 / 0.69						
RoBERTa	0.81 / 0.84	0.77 / 0.69	0.81 / 0.71						
Metaphor RoBERTa	0.83 / 0.85	0.80 / 0.74	0.82 / 0.71						

Baseline Results Across Tasks. Genre pair differences are also notable. The *Poetry vs. Novel* task is consistently the easiest across all models and languages, with average F1 scores generally exceeding 0.80. Poetry tends to employ distinctive linguistic features that are less common in prose forms. By contrast, both novels and dramas belong broadly to prose traditions: novels typically feature extended narrative, descriptive exposition, and inner monologue, whereas drama is conveyed primarily through dialogue and stage directions, but both share similar syntactic and lexical patterns. Consequently, it's easier for models to distinguish poetry from prose, while novel and drama are more linguistically proximate and thus harder to differentiate computationally.

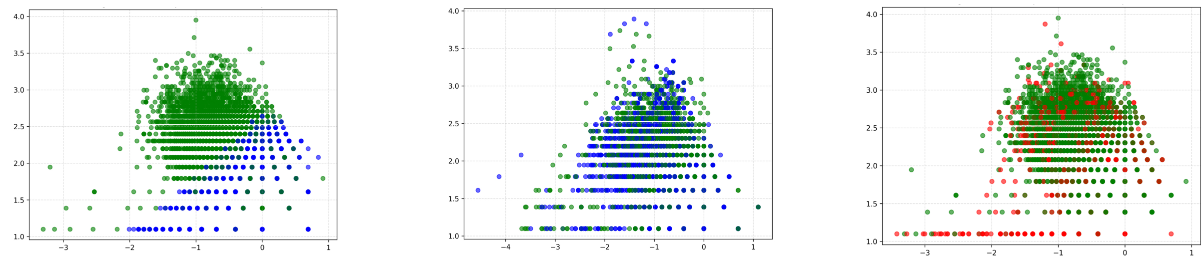
For the Novel vs. Drama task, all training methods fail to outperform the baseline, which is understandable given that these two genres often share similar narrative structures and linguistic features, making them inherently more difficult to distinguish. Beyond this, we observe varying degrees of improvement on the other two tasks across the four models.

Overall, syntax tree information and metaphor count do not appear to significantly enhance the models' ability to differentiate between genres. A more detailed analysis will be provided in Section 4.3.In contrast, the metre pattern feature proves to be the most robust and consistently beneficial across tasks and models. It delivers steady F1 score gains in both the Poetry vs. Novel and Poetry vs. Drama tasks, often outperforming other features by a margin of 2–7%. These results underscore the importance of prosodic and rhythmic cues as fundamental signals of genre—particularly in poetry, where metrical structure is deeply embedded in the form.

Baseline Results Across Models. Across all languages and tasks, Metaphor-RoBERTa consistently demonstrates strong performance. It achieves the highest average F1 scores in both the *Poetry vs. Novel* (0.83, 0.85) and *Poetry vs. Drama* (0.80, 0.74) settings, and ties for the best result in *Novel vs. Drama* (0.82, 0.71). This suggests that metaphor-aware pretraining may help capture stylistic distinctions between literary genres, particularly in tasks involving poetry.

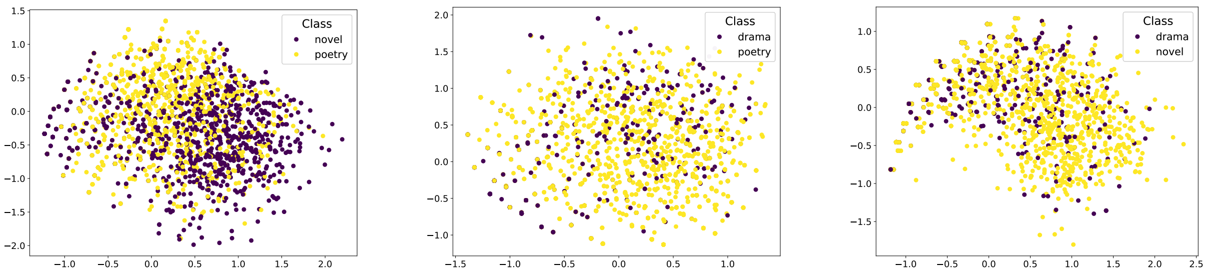
RoBERTa also performs robustly, particularly on *Poetry vs. Novel* tasks (0.81, 0.84), slightly trailing Metaphor-RoBERTa. Its advantage is especially pronounced in Romance languages such as Italian and Portuguese. DistilBERT, despite its smaller size, maintains competitive performance throughout. Notably, it outperforms BERT in many cases, including the French *Poetry vs. Novel* task (0.76, 0.59 vs. 0.50, 0.40), demonstrating that reduced model capacity does not significantly compromise genre classification when fine-tuned properly.

Model-wise, BERT benefits the most across all three linguistic features, likely due to its larger capacity to leverage explicit signals. DistilBERT shows improvements primarily with metre patterns, possibly compensating for its reduced model size with clearer prosodic cues. RoBERTa and Metaphor-RoBERTa demonstrate mixed results, with metaphor-based fine-tuning occasionally lessening the positive impact of added linguistic fea-



(a) poetry vs. novel (b) poetry vs. novel (French) (c) novel vs. drama

Figure 3: Syntactic tree analysis. The x-axis represents $\log(\text{depth\_ratio})$, while the y-axis represents $\log(\text{tree\_depth}+1)$. Green dots indicate novels, blue denote poetry, and red represent drama. Subfigure (a) shows the plot for Poetry vs. Novel in English, which is clearly linearly separable. Subfigure (b) presents the Poetry vs. Novel contrast in French, revealing a more complex distribution. Subfigure (c) displays the Novel vs. Drama set in English, which also exhibits significant overlap and is difficult to separate.



(a) poetry vs. novel (b) drama vs. novel (c) drama vs. poetry

Figure 4: Metre pattern analysis. We extract metre patterns from the raw texts and represent them as binary feature vectors, where each bit corresponds to a rhythmic unit. These vectors are padded to a uniform length and projected into a two-dimensional latent space using Principal Component Analysis (PCA).

tures, possibly due to feature redundancy or overfitting.

4.3 Results Analysis

In this section, we analyze how three linguistic features contribute to improving model performance on the classification task.

Syntax Tree Information. As shown in Table 3, *Metre Pattern* achieved the most substantial performance improvement among the three linguistic features. The other two features yielded only slight gains, primarily on the *Poetry vs. Novel* classification task, and showed limited effectiveness on the other two genre pairs. This can be explained from a linguistic perspective. Taking English as an example, Figure 3 reveal that syntax tree depth and depth ratio exhibit clear linear separability between poetry and novel, whereas the distinction between drama and novel is far less apparent. Given the relatively distinct boundary between poetry and novel, it is likely that the model is already capable of capturing syntactic information directly from raw sentences, which may explain the modest gains observed when these syntactic features are added through our fine-tuning method. Additionally, we use the same metric to evaluate the French dataset. The results reveal that French poetry and novels are not easily separable based on syntactic tree information, which may also explain the limited improvement observed on the French data.

Metaphor Count. Meanwhile, we calculated the average number of metaphorical words per sentence in English texts across three genres: drama (1.38), novel (2.50), and poetry (1.39). The results show that drama and poetry share similar levels of metaphor usage, while novels tend to contain significantly more metaphorical expressions on average. However, since the overall differences are relatively subtle, this variation does not appear to strongly influence the model's ability to distinguish between genres. Although it does not directly contribute to performance, this method plays a key role in controlling for the potential impact of increased token length on our tasks. The results suggest that linguistic features such as metre patterns contribute

Table 3: Comparative Analysis of Feature Optimization Methods (F1 Score Changes)

Feature	Set	BERT		DistilBERT		RoBERTa		Metaphor RoBERTa	
Syntax Tree Depth									
	Poetry+Novel	0.77 +1%	0.81 +5%	0.77 -3%	0.80 +2%	0.82 +1%	0.85 +1%	0.83 –	0.86 +1%
	Poetry+Drama	0.79 –	0.72 -1%	0.78 -1%	0.72 -1%	0.77 –	0.72 +3%	0.79 -1%	0.66 -8%
	Novel+Drama	0.79 –	0.68 -1%	0.78 -1%	0.69 –	0.81 –	0.71 –	0.80 -2%	0.67 -4%
Metaphor Count									
	Poetry+Novel	0.78 +2%	0.80 +4%	0.78 -2%	0.80 +2%	0.82 +1%	0.85 +1%	0.84 +1%	0.86 +1%
	Poetry+Drama	0.79 –	0.73 –	0.78 -1%	0.73 –	0.77 –	0.71 +2%	0.80 -2%	0.76 +2%
	Novel+Drama	0.78 -1%	0.69 –	0.79 –	0.69 –	0.81 –	0.69 -2%	0.81 -1%	0.70 -1%
Metre Pattern									
	Poetry+Novel	0.80 +4%	0.83 +7%	0.79 -1%	0.81 +3%	0.79 -2%	0.84 –	0.83 –	0.87 +2%
	Poetry+Drama	0.80 +1%	0.72 -1%	0.79 –	0.74 +1%	0.78 +1%	0.73 +4%	0.81 +1%	0.76 +2%
	Novel+Drama	0.79 –	0.69 –	0.78 -1%	0.69 –	0.79 -2%	0.66 -5%	0.77 -5%	0.66 -5%

Color of notation: ↑ Improvement ↓ Decline **Color of cell**: Improvement ≥ 0.02 | Decline ≥ 0.02 (non-negligible changes)

to performance improvements that go beyond the effects of token length, highlighting the intrinsic value of the features themselves.

Metre Pattern. Most importantly, the average F1 score has increased significantly when metre patterns are taken into consideration. This suggests that metre patterns encode subtle but meaningful information that improves genre classification performance, even if such patterns are not easily separable in low-dimensional space. As shown in Figure 4, the PCA projection of the binary metre vectors doesn't clearly clustered by class, indicating that the informative features may be distributed across multiple dimensions and not linearly separable. This reinforces the idea that metre patterns contribute in a complex, high-dimensional manner that benefits classification models, even if they are not directly interpretable via 2D visualization. The improved F1 scores highlight that models can detect and leverage these latent rhythmic signals effectively, even when they are not visually obvious.

Limitations

While our study introduces a linguistically enriched, multilingual dataset and provides novel insights into literary genre classification with LLMs, several limitations remain. First, the dataset is still constrained by the availability and biases of Project Gutenberg texts. These works tend to overrepresent canonical literature from specific historical periods and underrepresent contemporary, non-Western, or marginalized voices, which may limit the generalizability of our findings. Second, the extraction of linguistic features such as metaphor and phonetic regularity relies on heuristic or proxy-based methods, which may miss nuanced or culturally specific expressions. Future work should explore the inclusion of more diverse corpora and refine feature extraction with neural or hybrid methods to support interpretability across typologically diverse languages.

Conclusion

This study presents a novel multilingual dataset for literary genre classification, enhanced with linguistically informed features such as syntactic depth, metaphor frequency, and metrical patterns. By integrating these cues into large language model pipelines, we show that genre distinctions can be computationally captured through measurable linguistic signals. Our results indicate that these features improve genre discrimination across languages, offering both performance gains and interpretability. This work lays the groundwork for future research at the intersection of linguistics, literary analysis, and AI, supporting cross-cultural stylistic inquiry and deeper engagement with the structural and figurative dimensions of text. The future direction of this research involves expanding the dataset to cover more languages and literary traditions, enabling broader cross-cultural comparisons, and applying the method to broader and more complex tasks.

Ethical Considerations

This work relies exclusively on publicly available, public-domain texts from Project Gutenberg, ensuring that no copyrighted or private data are used. As a result, the study does not raise concerns related to personal privacy, consent, or data ownership. However, we acknowledge that Project Gutenberg predominantly represents canonical literary works from specific historical periods and cultural traditions, which may introduce biases toward Western, Eurocentric, and historically privileged literary forms. In addition, some linguistic features used in this study—such as metaphor detection and metrical analysis—are derived from automated tools and pretrained models, which may encode their own biases or inaccuracies. Finally, we emphasize that our findings are intended to support linguistic and literary analysis, rather than to make normative judgments about literary value or cultural significance.

References

Carter Anderson, Mien Nguyen, and Rolando Coto-Solano. 2025. Unsupervised, semi-supervised and llm-based morphological segmentation for bribri. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 63–76.

Ehsaneddin Asgari, Yassine El Kheir, and Mohammad Ali Sadraei Javaheri. 2025. Morphbpe: A morpho-aware tokenizer bridging linguistic complexity for efficient llm training across morphologies. *arXiv preprint arXiv:2502.00894*.

Nicolas Ballier, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès. 2023. Using whisper llm for automatic phonetic diagnosis of l2 speech: A case study with french learners of english. In *6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, volume 6.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Towards llm-guided causal explainability for black-box text classifiers. *Preprint*, arXiv:2309.13340.

Ivan Brigadoi. 2021. Genre classification using syntactic features.

Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. 2024. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234.

Noam Chomsky. 2002. *Syntactic structures*. Mouton de Gruyter.

Pietro Dell'Oglio, Dominique Brunato, and Felice Dell'Orletta. 2018. Lexicon and syntax: Complexity across genres and language varieties. In *CLiC-it*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Jaiv Doshi, Nathan Sparks, and Daniel Youn. Phonetic representation for language modeling with rhyme.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Sanda M Harabagiu, Dan I Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. Falcon: Boosting knowledge for answer engines. In *TREC*, volume 9, pages 479–488.

Rebecca M. M. Hicke and David Mimno. 2025. Looking for the inner music: Probing llms' understanding of literary style. *Preprint*, arXiv:2502.03647.

Gueyoung Jung, Matti A Hiltunen, Kaustubh R Joshi, Richard D Schlichting, and Calton Pu. 2010. Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures. In *2010 IEEE 30th International Conference on Distributed Computing Systems*, pages 62–73. IEEE.

Mikhail V Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. 2024. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models. *arXiv preprint arXiv:2410.22660*.

Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction*, 5(3):1149–1175.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2022. The ginco training dataset for web genre identification of documents out in the wild. *Preprint*, arXiv:2201.03857.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, page 145–150, New York, NY, USA. Association for Computing Machinery.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, and 1 others. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikola Ljubešić and Taja Kuzman. 2024. Classla-web: Comparable web corpora of south slavic languages enriched with linguistic and genre annotation. *Preprint*, arXiv:2403.12721.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Frederick J. Newmeyer. 2023. *Chomsky and the Turn to Syntax, Including Alternative Approaches to Syntax*, page 549–576. Cambridge University Press.

Shinwoo Park, Shubin Kim, Do-Kyung Kim, and Yo-Sub Han. 2025. Katfishnet: Detecting llm-generated korean text through linguistic feature analysis. *Preprint*, arXiv:2503.00032.

Ariel Rosenfeld and Teddy Lazebnik. 2024. Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard. *arXiv preprint arXiv:2402.14533*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Serge Sharoff. 2018. Functional text dimensions for the annotation of web corpora. *Corpora*, 13(1):65–95.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lennart Wachowiak, Dagmar Gromann, and Chao Xu. 2022. Drum up support: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, 29(4):534–561.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Viktor Maksimovic Žirmunskij. 2016. *Introduction to metrics: the theory of verse*, volume 58. Walter de Gruyter GmbH & Co KG.

A Detailed Performance under Different Linguistic Features

Below, we present the detailed F1 scores obtained under our three linguistic feature–based training approaches. The experimental settings are kept consistent with those used in the baseline model to ensure fair comparison. The only difference lies in the incorporation of linguistic features—each model is enhanced by adding one of the following types of features: syntactic (e.g., syntax tree depth), phonological (e.g., metre patterns), or stylistic (e.g., metaphor usage). These features are intended to inject deeper linguistic knowledge into the model, thereby improving its ability to distinguish between genres. The results provide insight into the individual contributions of these features to classification performance.

Table 4: F1 scores by genre and language with **syntax tree depth** considered.

Language	Set	BERT	DistilBERT	RoBERTa	Metaphor RoBERTa
EN	Poetry + Novel	0.97+0.96	0.96+0.95	0.97+0.96	0.97+0.96
	Poetry + Drama	0.97+0.88	0.96+0.86	0.97+0.89	0.96+0.87
	Novel + Drama	0.90+0.68	0.90+0.67	0.90+0.64	0.89+0.62
FR	Poetry + Novel	0.64+0.74	0.66+0.71	0.78+0.82	0.75+0.81
	Poetry + Drama	0.72+0.72	0.77+0.73	0.69+0.69	0.73+0.70
	Novel + Drama	0.72+0.70	0.71+0.65	0.75+0.74	0.76+0.70
DE	Poetry + Novel	0.73+0.77	0.72+0.77	0.74+0.76	0.80+0.83
	Poetry + Drama	0.76+0.67	0.74+0.69	0.75+0.70	0.76+0.74
	Novel + Drama	0.78+0.75	0.78+0.72	0.81+0.74	0.80+0.74
ES	Poetry + Novel	0.75+0.77	0.77+0.76	0.79+0.81	0.78+0.81
	Poetry + Drama	0.76+0.69	0.70+0.69	0.72+0.71	0.69+0.22
	Novel + Drama	0.74+0.69	0.74+0.71	0.76+0.67	0.79+0.64
IT	Poetry + Novel	0.76+0.79	0.74+0.80	0.77+0.84	0.82+0.85
	Poetry + Drama	0.78+0.72	0.79+0.73	0.75+0.71	0.80+0.77
	Novel + Drama	0.79+0.69	0.80+0.72	0.83+0.74	0.81+0.71
PT	Poetry + Novel	0.76+0.81	0.76+0.81	0.86+0.89	0.86+0.89
	Poetry + Drama	0.73+0.62	0.74+0.65	0.71+0.64	0.77+0.67
	Novel + Drama	0.78+0.60	0.76+0.68	0.80+0.70	0.73+0.62
Average	Poetry + Novel	0.77+0.81	0.77+0.80	0.82+0.85	0.83+0.86
	Poetry + Drama	0.79+0.72	0.78+0.72	0.77+0.72	0.79+0.66
	Novel + Drama	0.79+0.68	0.78+0.69	0.81+0.71	0.80+0.67

Table 5: F1 scores by genre and language with **metaphor count** considered.

Language	Set	BERT	DistilBERT	RoBERTa	Metaphor RoBERTa
EN	Poetry + Novel	0.97+0.96	0.97+0.97	0.97+0.96	0.96+0.96
	Poetry + Drama	0.97+0.89	0.97+0.90	0.97+0.88	0.97+0.89
	Novel + Drama	0.91+0.69	0.91+0.68	0.89+0.59	0.90+0.64
FR	Poetry + Novel	0.66+0.73	0.70+0.67	0.74+0.82	0.79+0.81
	Poetry + Drama	0.75+0.73	0.71+0.73	0.69+0.70	0.64+0.72
	Novel + Drama	0.71+0.72	0.72+0.68	0.76+0.70	0.74+0.68
DE	Poetry + Novel	0.73+0.75	0.73+0.78	0.80+0.84	0.80+0.84
	Poetry + Drama	0.76+0.69	0.72+0.71	0.73+0.70	0.76+0.73
	Novel + Drama	0.78+0.73	0.78+0.73	0.81+0.76	0.79+0.76
ES	Poetry + Novel	0.76+0.74	0.79+0.77	0.79+0.76	0.81+0.82
	Poetry + Drama	0.74+0.73	0.73+0.72	0.74+0.71	0.77+0.75
	Novel + Drama	0.71+0.68	0.77+0.67	0.80+0.70	0.81+0.68
IT	Poetry + Novel	0.77+0.79	0.76+0.81	0.75+0.83	0.79+0.85
	Poetry + Drama	0.78+0.73	0.79+0.73	0.75+0.70	0.80+0.79
	Novel + Drama	0.79+0.69	0.81+0.69	0.83+0.73	0.81+0.72
PT	Poetry + Novel	0.79+0.82	0.75+0.80	0.86+0.89	0.87+0.89
	Poetry + Drama	0.73+0.63	0.73+0.60	0.75+0.57	0.75+0.67
	Novel + Drama	0.79+0.66	0.77+0.67	0.78+0.64	0.82+0.71
Average	Poetry + Novel	0.78+0.80	0.78+0.80	0.82+0.85	0.84+0.86
	Poetry + Drama	0.79+0.73	0.78+0.73	0.77+0.71	0.78+0.76
	Novel + Drama	0.78+0.69	0.79+0.69	0.81+0.69	0.81+0.70

Table 6: F1 scores by genre and language with **metre pattern** considered.

Language	Set	BERT	DistilBERT	RoBERTa	Metaphor RoBERTa
EN	Poetry + Novel	0.98+0.97	0.98+0.97	0.97+0.97	0.97+0.97
	Poetry + Drama	0.97+0.90	0.97+0.91	0.97+0.89	0.97+0.90
	Novel + Drama	0.91+0.68	0.89+0.68	0.90+0.58	0.87+0.58
FR	Poetry + Novel	0.71+0.74	0.75+0.69	0.78+0.83	0.76+0.82
	Poetry + Drama	0.74+0.70	0.77+0.73	0.73+0.67	0.77+0.73
	Novel + Drama	0.74+0.69	0.72+0.70	0.75+0.73	0.74+0.73
DE	Poetry + Novel	0.73+0.79	0.74+0.77	0.79+0.83	0.80+0.85
	Poetry + Drama	0.75+0.70	0.74+0.70	0.73+0.72	0.76+0.73
	Novel + Drama	0.78+0.73	0.76+0.73	0.80+0.76	0.70+0.66
ES	Poetry + Novel	0.79+0.81	0.78+0.77	0.80+0.82	0.80+0.82
	Poetry + Drama	0.77+0.70	0.75+0.71	0.72+0.69	0.75+0.74
	Novel + Drama	0.76+0.68	0.76+0.66	0.66+0.54	0.80+0.70
IT	Poetry + Novel	0.80+0.83	0.72+0.80	0.75+0.83	0.81+0.85
	Poetry + Drama	0.79+0.75	0.79+0.75	0.80+0.79	0.82+0.78
	Novel + Drama	0.79+0.71	0.80+0.74	0.82+0.67	0.81+0.69
PT	Poetry + Novel	0.81+0.85	0.80+0.84	0.62+0.78	0.86+0.89
	Poetry + Drama	0.76+0.60	0.73+0.63	0.73+0.62	0.76+0.70
	Novel + Drama	0.78+0.62	0.77+0.63	0.78+0.68	0.70+0.63
Average	Poetry + Novel	0.80+0.83	0.79+0.81	0.79+0.84	0.83+0.87
	Poetry + Drama	0.80+0.72	0.79+0.74	0.78+0.73	0.81+0.76
	Novel + Drama	0.79+0.69	0.78+0.69	0.79+0.66	0.77+0.66