electronics

MDPI

Article

Comparative Evaluation of Transformer-Based Models for Plain Language Classification in Hungarian Legal–Administrative Texts

István Üveges

Griffsoft Zrt., 1041 Budapest, Hungary; istvan.uveges@griffsoft.hu

Abstract

Plain Language seeks to enhance the clarity and comprehensibility of legal and administrative communication; while Natural Language Processing (NLP) offers promising tools for assessing text complexity, most Plain Language classification studies focus exclusively on English, leaving low-resource languages underexplored. This study presents the first systematic evaluation of transformer-based models for sentence-level Plain Language classification in Hungarian tax administrative texts. We benchmarked zero-shot prompting with GPT-4o against fine-tuned open-weight and proprietary models, including huBERT, XLM-RoBERTa, GPT-4o-mini, and Gemini 1.0 Pro, and contextualized these results against previously established lightweight machine learning baselines based on term frequency-inverse document frequency with a support vector machine (TF-IDF + SVM) and fastText. To address data scarcity, we applied translation-based data augmentation using parallel Hungarian–English corpora. The best-performing model achieved a macro-average F1-score of 0.79. Mid-sized models also delivered competitive results, combining accuracy with feasible inference speed and deployment flexibility. Beyond classification performance, we conducted local and aggregated interpretability analysis based on Shapley-values to identify linguistic patterns influencing model decisions. This revealed alignment with known Plain Language features, such as nominalizations and syntactic complexity, as well as biases introduced by frequent domain-specific terms. Our findings demonstrate that Plain Language classifiers can be effectively adapted to low-resource legal–administrative domains. The results support the development of real-time feedback tools that promote linguistic accessibility and contribute to the broader goal of Access to Justice.

Keywords: Plain Language; low-resource languages; transformer models; data augmentation; legal–administrative text classification

Academic Editors: Mayra Antonio-Cruz and Carlos Alejandro Merlo-Zapata

Received: 30 May 2026
Revised: 20 June 2026
Accepted: 26 June 2026
Published: 6 July 2026

1. Introduction

As a result of accelerated technological development, the global trend of digitization, and the growing complexity of legal regulations, the quality of public communication by state bodies has perhaps never been more crucial than it is today. This is especially true for legal and administrative texts that help individuals understand their obligations and enforce their rights.

The quality of such communication can be evaluated from multiple perspectives, including clarity, timeliness, accessibility, and trustworthiness, all of which are key elements of effective government communication during crises [1]. However, these factors are no less

important outside of emergency contexts. In fact, they are essential for enabling citizens to navigate everyday bureaucratic processes and to exercise their legal rights.

One of the most significant obstacles to this goal is the linguistic complexity of legal and administrative texts, which often creates barriers to understanding. In the absence of clear and accessible language, such documents can be misinterpreted, turning a communication issue into a matter of social justice. Improving the clarity of legal texts is therefore closely tied to the principle of Access to Justice [2–4], a fundamental requirement of the rule of law [5,6].

Access to Justice depends not only on the availability and reliability of information but also on whether people can understand official documents addressed to them, and how quickly they can do so [7]. These documents include, for example, informational materials from public offices and the texts of laws and regulations.

This concern has given rise to the Plain Language Movement (PLM), which has advocated since the 1970s for clearer public and legal communication [8]. According to the U.S. government's official guidelines, "a communication is in Plain Language if its wording, structure, and design are so clear that the intended audience can easily find what they need, understand what they find, and use that information" (https://www.plainlanguage.gov/guidelines/ (accessed on 30 May 2026)). The primary aim of the movement is to make public and legal documents more comprehensible, thereby helping citizens understand their rights and obligations.

Plain Language efforts have inspired a range of national and international initiatives, including the ISO 24495-1 standard [9] (https://www.iso.org/standard/78907.html (accessed on 30 May 2026)), which sets out formal criteria for clarity and accessibility. The development of this standard is coordinated by the International Plain Language Federation, an umbrella organization that includes Clarity International and the Plain Language Association International (https://www.iplfederation.org/iso-standard/ (accessed on 30 May 2026)).

Although awareness of Plain Language (PL) has grown significantly in recent years, legal documents often remain linguistically complex. Long sentences, archaic terminology, and nested clauses continue to hinder comprehension. Revising texts that were not originally written in PL is particularly challenging, as it demands both specialized expertise and substantial time. This is especially true in the legal and administrative domain, where precision and clarity must coexist, making manual simplification both labor-intensive and costly. To address this, Natural Language Processing (NLP) techniques have gained increasing relevance. They offer the potential for semi-automated simplification, improving efficiency and enhancing access to legal information.

While PL-oriented tools and initiatives have made considerable progress internationally, they have mostly focused on English-language texts. Advanced clarity testers (e.g., Readability Consensus, Flesch–Kincaid) and artificial intelligence (AI)-driven writing assistants (e.g., Grammarly, Hemingway Editor) are readily available in English, combining rule-based and machine learning approaches to identify and improve complex sentence structures.

In contrast, smaller languages like Hungarian remain under-served in this area. Foundational NLP tools such as HuSpaCy [10] support basic linguistic processing, but do not offer capabilities for simplification or paraphrasing. Furthermore, the only publicly available Hungarian simplification corpus, huPWKP [11], is based on the English Parallel Wikipedia Simplification Corpus [12] and reflects the principles of Simple English rather than the stricter clarity and accessibility requirements needed for legal and administrative communication.

To address these limitations, another Hungarian-language corpus specifically designed for PL processing has recently been introduced [13]. This dataset contains parallel versions

of legal and administrative documents from the domain of tax administration in their original form and their PL reformulations, created by linguistic experts. The parallel structure provides a unique opportunity for machine-learning-based adaptation, enabling models to learn how complex, technical language can be transformed into clearer and more accessible text. Preliminary experiments using this corpus have already demonstrated its potential for automated classification and simplification tasks [14].

Nevertheless, the development of robust machine learning models for PL adaptation in Hungarian legal–administrative texts remains an open research challenge; while large language models (LLMs) such as GPT-4 have shown considerable success in generating PL summaries in English (especially in structured domains like biomedical literature [15]), their effectiveness in legal contexts is less established. In particular, the need to preserve precise normative content while enhancing clarity introduces unique difficulties that general-purpose models are not yet fully equipped to resolve.

This technological gap underscores the need for localized solutions that can support PL adaptation in non-English contexts. Although international efforts have advanced PL processing in major languages, domain-specific tools for underrepresented languages like Hungarian are still largely missing. Bridging this gap is essential to ensure that the benefits of Plain Language, enhanced comprehensibility, efficiency, and Access to Justice, are available across linguistic boundaries.

In response to this challenge, the present study aims to develop a machine-learned component for a Hungarian legal–administrative writing assistant. This component is designed to automatically classify each sentence as either "Plain Language"or "not Plain Language", thereby providing real-time feedback during drafting and revision. Such functionality can significantly reduce the time and expertise required for manual simplification, while supporting legal professionals in producing more accessible documents.

To this end, we investigate the potential of sentence-level PL classification through fine-tuning several state-of-the-art transformer-based models, including XLM-RoBERTa, the Bidirectional Encoder Representations from Transformers (BERT) model, Gemini 1.0 Pro, and GPT-4o-mini. These models are trained and evaluated on Hungarian legal and administrative corpora that contain parallel PL annotations, enabling supervised learning. Our approach not only assesses the binary classification performance of each model but also examines their ability to identify complex sentence structures that are likely to require simplification.

This study is methodologically focused on sentence-level classification, rather than generative rewriting or translation-based simplification. We argue that this approach is more feasible for low-resource settings and legally sensitive domains, where preserving meaning and minimizing distortion are paramount.

In addition to benchmarking model performance, we also address the interpretability of classification decisions using SHAP (SHapley Additive Explanations) analysis [16]. By visualizing the contribution of individual tokens to the model's predictions, we provide linguistic insight into how legal complexity is operationalized in transformer-based classifiers. This supports both error analysis and the development of transparent, domain-aware NLP applications in legal settings.

Although the empirical focus of this study is on Hungarian legal texts, the proposed methods and challenges addressed here are applicable to other under-resourced languages and domains that require precision-driven simplification.

The main contributions of this study are therefore:

1. We motivate the use of sentence-level classification as a viable alternative to machine translation, particularly in contexts where parallel PL corpora are limited in size.

2. We release and evaluate transformer-based PL classifiers optimized for integration into a future Hungarian legal writing assistant.

The remainder of this paper is structured as follows. Section 2 provides a review of the relevant literature on Plain Language and Text Simplification (TS). Section 3 reviews machine learning approaches to TS. Section 4 describes the dataset and preprocessing procedures. Section 5 presents the methodological framework. Section 6 reports the experimental results, followed by a discussion in Section 7. Finally, Section 8 concludes the paper, and outlines directions for future work.

2. Literature Review

Research on Plain Language (PL) spans several disciplines, including linguistics, legal studies, communication theory, and, more recently, computer science. This section reviews the major theoretical and empirical contributions that have shaped the field. It pays particular attention to early readability measures, psycholinguistic insights, legal reform movements, and the unique challenges faced in the Hungarian context. The goal is to situate recent NLP-based developments within a broader historical and interdisciplinary framework.

2.1. Historical Background and the Emergence of Plain Language

The concept of PL emerged in the 1970s as part of a broader push for transparency and accessibility in legal and governmental communication [8]. This movement was motivated by concerns over the exclusionary nature of bureaucratic and legal language. Since then, PL has become a key component of legal reform efforts aiming to reduce linguistic complexity and foster civic participation [17].

Major milestones include the U.S. Plain Writing Act of 2010 and the Federal Plain Language Guidelines of 2011. The internationalization of these efforts culminated in the ISO 24495-1:2023 standard [9], developed by the International Plain Language Federation. This global benchmark codifies essential PL principles such as clarity, accessibility, and usability for both digital and print media.

2.2. Readability Formulas and Early Stage Approaches

One of the first quantitative approaches to TS involved readability formulas. Originating in the early 20th century [18,19], these metrics aimed to assess text difficulty using surface features such as sentence length and word frequency. The Flesch Reading Ease, FOG Index, Fry Graph, and Dale-Chall formula became widely used in educational and administrative settings.

These methods gained institutional prominence when adopted by the U.S. military in the 1950s [20]. Their legacy persists in modern tools such as Microsoft Word's readability checker [21]. However, their reliance on shallow linguistic indicators limits their ability to capture syntactic or semantic complexity [22,23]. In addition, they fail to account for discourse-level cohesion and domain-specific terminology.

The distinction between "readability" and "legibility" also emerged during this period, while readability focuses on linguistic features, legibility encompasses visual presentation, layout, and typographic choices [24]. Dale and Chall's expanded definition emphasized not only ease of decoding but also reader engagement and comprehension success [18].

2.3. Psycholinguistics and Comprehensibility Metrics

Psycholinguistic approaches shifted attention from text-internal features to cognitive processing. This research examines how readers comprehend syntactic structures, interpret vocabulary, and resolve ambiguities [25–27]. Empirical methods include eye-tracking,

reaction time analysis, and neuroimaging techniques such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) [28].

Notably, Charrow and Charrow's study of jury instructions revealed that legal language, characterized by embedded clauses, passive constructions, and archaic terms, poses significant barriers to lay understanding [29]. These findings have directly informed the design of PL principles and legislative communication strategies.

In the Hungarian context, psycholinguistic research has identified syntactic structures that increase processing difficulty. These include preverbal noun pileups, the separation of verbs from their prefixes, and deviations from canonical word order [28,30,31]. Such patterns are particularly relevant in legal writing and underscore the need for language-sensitive simplification tools.

2.4. Plain Language Movement and Legal Language Studies

The PLM has been conceptualized not only as a communication reform, but also as a form of bottom-up language planning and critique. It arose in response to public dissatisfaction with opaque official language, particularly in English-speaking countries [32–34].

Key developments include the U.S. federal mandates for plain writing and widely used guides for legal professionals [35–38]. The ISO 24495-1 standard now formalizes these efforts into a global framework, offering practical benchmarks for document design across sectors.

2.5. Legal Language and Comprehensibility in the Hungarian Context

In Hungary, empirical studies have highlighted the impact of linguistic opacity on legal fairness and procedural transparency [7]. Corpus-based analyses have revealed widespread use of nominalizations, legalese, and syntactic discontinuities in judicial and administrative texts [39–41].

Parallel research within the "law and language" tradition has investigated the linguistic dimensions of courtroom interaction, freedom of speech, and the pragmatic use of legal terms [42–44]. Hungarian scholars have recently extended these inquiries using corpus linguistics to map comprehensibility-related features in legislative texts [45,46]. Practice-oriented Hungarian initiatives have likewise sought to make public-law concepts accessible to non-specialist readers while preserving their normative content, combining expert review with lay-reader testing [47].

These findings provide critical groundwork for domain-specific NLP solutions that support Access to Justice through improved legal communication.

3. Machine Learning and Text Simplification

As PL has gained increasing attention in legal and administrative communication, NLP techniques have emerged as essential tools to support its implementation. This section reviews the growing body of research on TS in the field of NLP, with a particular focus on the distinction between Easy English and PL, the framing of simplification tasks, and the application of Machine Learning (ML) and Large Language Models (LLMs) to these challenges. We emphasize the importance of distinguishing between simplification approaches that prioritize accessibility versus those that maintain legal precision, and we identify key limitations in applying existing TS models to domain-specific PL tasks.

3.1. Easy English vs. Plain English

TS is a well-established subfield of NLP that aims to make written content more accessible without altering its original meaning [48,49]. However, there is an important distinction between Easy English and PL, which is often blurred in ML-driven simplification efforts.

Easy English, often represented by datasets like Simple Wikipedia, is designed for individuals with limited literacy, non-native speakers, or those with cognitive impairments. It prioritizes basic vocabulary and simplified syntactic structures, often at the expense of semantic richness and precision [50]. For example, Easy English texts are typically constrained to the most common 1000 words in English, making them accessible but less suitable for legal or administrative texts where specific terminology is required.

In contrast, PL aims to make texts more understandable for the average reader without compromising on accuracy or legal integrity. It targets unnecessary linguistic complexity while ensuring that the original legal or administrative meaning remains intact. In German linguistic literature, this distinction is marked by the terms *Leichte Sprache* (Easy Language) and *Einfache Sprache* (Plain Language) [51,52]; while *Leichte Sprache* is intended for those with cognitive impairments, *Einfache Sprache* focuses on enhancing clarity for general readers without simplifying legal content to the point of distortion.

The majority of TS research has historically focused on Easy English due to the availability of large parallel corpora such as Simple Wikipedia. These resources have enabled ML models to learn patterns of simplification, but they have also skewed research towards a form of simplification that does not meet the criteria of PL. PL, particularly in legal and administrative contexts, requires more nuanced approaches that balance clarity with legal accuracy; a challenge not addressed by most existing TS models.

However, recent studies have begun to bridge this gap. For instance, the Newsela Corpus [53] includes graded versions of news articles adapted for different reading levels, which partially aligns with PL principles. Similarly, ASSET [54] extends the concept of simplification by introducing multiple rewritten versions of the same sentence, covering various levels of complexity reduction. Despite these advances, there is still a substantial gap in resources explicitly designed for PL transformation, especially in highly regulated fields like law and public administration.

In Hungary, the huPWKP corpus represents the only significant effort towards simplification, although it is based on Simple Wikipedia and does not meet Plain Language standards [11]. By contrast, recent work on Hungarian tax administration texts introduces a legal-oriented simplification corpus with parallel Plain Language versions, which is much closer to PL adaptation [13,14,55].

The distinction between Easy English and Plain Language is crucial for legal and administrative texts, where legal precision cannot be compromised for simplicity. Therefore, the choice of approach significantly impacts the effectiveness and applicability of automated simplification tools in these domains.

3.2. Task Framing: Classification vs. Translation

While many TS systems frame simplification as a monolingual machine translation (MT) problem, such models typically require extensive parallel corpora. For Hungarian legal texts, such resources are extremely limited. As a result, our study follows an alternative approach: we cast the task as binary sentence-level classification (Plain vs. Non-Plain), which is more suitable for low-resource and high-precision domains. This reframing acknowledges the need for minimal distortion in legal texts and shifts the focus from full sentence rewriting to the detection of complex segments that may require revision.

3.3. Related Work in Machine Learning for Text Simplification

ML-based TS methods began as rule-based systems, which applied handcrafted syntactic and lexical rules to convert complex sentences into simpler versions [56,57]. Although interpretable, such methods were brittle and lacked adaptability to diverse linguistic contexts.

The development of data-driven approaches allowed models to learn simplification patterns from parallel corpora. EditNTS [58], for instance, uses a neural programmer-interpreter model to rewrite sentences. Modern neural architectures, including Seq2Seq [59], Transformer [60], and pretrained language models like BERT [61] and T5 [62], have shown impressive performance in simplification tasks.

Recent efforts have demonstrated the value of domain-specific fine-tuning. In biomedical NLP, BART [63] and T5 [62] have been fine-tuned for clarity-preserving simplification. The PLABA dataset [64] showed that BART-Large with control tokens achieved the highest SARI score (46.54), while T5-base performed best on BERTScore (72.62). The Med-EASi project [65] introduced fine-grained control over linguistic transformations—lexical, syntactic, and structural—through transformer-based models.

Despite these successes, most benchmark datasets (e.g., WikiLarge [66]) focus on Easy English. Therefore, models trained on them risk producing outputs that lack the legal precision required for PL. Moreover, these datasets are not aligned with legal or administrative communication goals, reducing their relevance for domain-specific applications.

3.4. Large Language Models and Legal Domain Constraints

LLMs like GPT-4 represent a significant leap in generative language capabilities. In healthcare, LLMs have demonstrated the ability to produce summaries that closely approximate expert-written PL content [15,67,68]. However, their application in law is limited by risks of semantic drift, normative distortion, and hallucinated content. Legal concepts may derive their meaning from a broader and carefully delimited doctrinal framework, and terms that appear closely related at the surface level may nevertheless refer to distinct legal categories [69]. Consequently, lexical simplification without sufficient domain knowledge may obscure conceptually relevant distinctions even when the resulting sentence remains grammatically fluent.

A recent evaluation of LLM-based simplification for Hungarian legal texts [70] revealed that despite their fluency, LLMs frequently alter legal meaning through subtle shifts: substituting disjunctions with conjunctions, misinterpreting modal verbs, or altering numeric conditions. Hallucinated legal references and inaccurate simplifications of terms like *előhaszonbérleti jog* [right of pre-emption] illustrate the difficulty of controlling outputs.

These findings highlight the importance of keeping humans in the loop. Until legal PL generation can guarantee semantic fidelity, hybrid workflows combining automated flagging with expert post-editing remain essential. Sentence-level classifiers offer a tractable and safe intermediary step (flagging complexity without rewriting) to support legal professionals in identifying passages that may require revision.

Table 1 summarizes the key TS models discussed and their relevance to PL-oriented applications.

Table 1. Comparison of major TS models and their applicability to PL tasks.

Model	Corpus	Domain	PL Suitability
BART-L w/CT	PLABA	Biomedical	Moderate–High
T5-base	PLABA, Med-EASi	Biomedical	Moderate
EditNTS	WikiLarge	Gen./Easy English	Low
GPT-4	Web + Finetune	Gen./Legal (exp.)	Variable/Risky
Sentence-level classifier	Custom PL corpora	Legal	High (Precise, Low Drift)

4. Data

This study leverages a curated Hungarian dataset specifically developed for PL adaptation in legal and administrative contexts. The dataset, previously introduced by [13],

contains parallel texts from the Hungarian National Tax and Customs Administration (Nemzeti Adó- és Vámhivatal, NAV). These texts were originally drafted in complex legalese and subsequently reformulated by linguistic experts to adhere to PL principles, with the aim of enhancing comprehensibility for the general public.

The corpus comprises various types of administrative communication, including informational brochures, guidance documents, and public service announcements. Each document appears in two distinct forms: the original, legally complex version, and a PL version rewritten for improved clarity and accessibility. Although these texts are not formal legislative acts, they are authoritative administrative communications that serve a quasi-legal function in informing citizens about rights, obligations, and procedures.

The original dataset consisted of 203 document pairs. During preprocessing, all texts were segmented into sentences and aligned at the sentence level. However, only sentence pairs that exhibited meaningful differences between the original and the PL version were retained. Sentences that appeared verbatim in both versions were excluded to ensure that the remaining corpus genuinely reflects stylistic and structural simplification.

The resulting corpus contains 5445 sentences from the Non-Plain documents and 5438 sentences from their PL counterparts. Each sentence was assigned a binary label based on its origin: "Non-Plain" if it appeared only in the original version, and "Plain" if it appeared only in the simplified version. These binary labels serve as the foundation for the sentence-level classification task described in the subsequent sections.

The statistics reported in Table 2 refer to the complete sentence-level corpus before partitioning. For model development and evaluation, the 203 document pairs were divided into training, validation, and test sets using an 80/10/10 ratio. The split was performed at the document-pair level rather than at the individual sentence level. Each pair consisted of an original administrative document and its corresponding Plain Language reformulation, and all sentences originating from the same document pair were assigned to the same subset. This procedure ensured that neither parallel sentence variants nor other sentences from the same source document could occur across the training, validation, and test sets.

Table 2. Basic statistics of the selected corpus (adapted from Table 1 of Ref. [13]).

	Non-Plain	**Plain**
Number of sentences	5445	5438
Number of tokens	139,704	127,945
Average sentence length	25.66	23.53
Percentage	50.03%	49.97%

This dataset represents a rare resource for Hungarian legal–administrative PL research and serves as a key contribution to low-resource language settings. It is particularly suitable for classification-based approaches, as its sentence-level structure provides clean, interpretable training examples without requiring large-scale generative modeling.

5. Methods

This section outlines the methodological framework employed to classify Hungarian legal texts based on their adherence to PL principles. Our approach combines zero-shot prompting, supervised fine-tuning of multiple pretrained language models, and data augmentation. Each component is detailed below, with a focus on their implementation for sentence-level classification in a low-resource, domain-specific context.

5.1. Zero-Shot Classification Setup

As a baseline, we employed the GPT-4o model [71] using a zero-shot prompting strategy. This approach allowed the model to classify legal texts without prior task-specific fine-tuning, relying entirely on natural language instructions. The primary goal was to evaluate the model's inherent understanding of PL principles in Hungarian texts, independent of extensive prompt engineering.

Zero-shot prompting is particularly relevant in low-resource language settings, such as Hungarian, where annotated domain-specific corpora are scarce [72], while LLMs show strong zero-shot performance in English [73], their effectiveness in smaller, underrepresented languages is limited [74]. Thus, zero-shot evaluation serves as a useful diagnostic tool to assess whether a model implicitly captures PL-related features in Hungarian.

To ensure task clarity and reduce output variability, we constructed a concise classification prompt:

> *You are a legal text classifier. Your task is to determine whether a legal text is written in Plain Language or not Plain Language. A text is considered 'Plain Language' if it is simple, clear, and easy to understand for a general audience without requiring legal expertise. Otherwise, it is 'not Plain Language'. Respond only with 'Plain Language' or 'not Plain Language'. Here is the text:*

This prompt was deliberately kept simple to avoid dependence on complex prompt-engineering strategies. It enabled deterministic, interpretable outputs that were benchmarked against supervised classifiers.

5.2. Fine-Tuning Setup

Fine-tuning involves adapting pretrained language models to a specific task through further training on labeled data. It enables the model to better capture domain-specific language use, making it especially valuable in low-resource and technical domains, such as legal text classification [75,76].

We fine-tuned four models selected for their relevance to multilingual and legal NLP tasks: huBERT, XLM-RoBERTa (XLM-R), GPT-4o-mini, and Gemini 1.0 Pro.

The huBERT model [77] is a Hungarian BERT-base model trained on a 9-billion-token Hungarian corpus. It contains 110M parameters and was fine-tuned for binary sentence classification. We used the Hugging Face Trainer API for training (max. 10 epochs, batch size 16, learning rate 5×10^{-6}). Tokenization used the pretrained huBERT tokenizer with a 512-token maximum sequence length. Optimization was performed with AdamW and early stopping after two stagnant epochs.

XLM-R [78] (base and large) was selected for its strong multilingual capabilities and cross-lingual alignment. Both variants were fine-tuned for a maximum of 10 epochs with a per-device training and evaluation batch size of 16, a learning rate of (5×10^{-5}), a maximum sequence length of 512 tokens, and a random seed of 42. The default AdamW optimizer of the Hugging Face Trainer was used. Evaluation and checkpoint saving were performed at the end of each epoch. Early stopping was applied with a patience of two consecutive evaluation cycles without improvement in validation loss. The checkpoint with the lowest validation loss was reloaded at the end of training, and only one checkpoint was retained.

OpenAI's GPT-4o-mini model [79] was fine-tuned via the OpenAI API using JSONL-formatted data. Prompts were sentence-level inputs as user prompts, paired with their corresponding labels (ie., Plain vs. Non-Plain). Default fine-tuning parameters were used, with convergence monitored through API-provided metrics.

Google's Gemini 1.0 Pro model [80] was fine-tuned via Google Cloud Vertex AI using the `gemini-1.0-pro-002` variant. Training occurred in the `europe-west2` region over

5 epochs with a learning rate multiplier of 1 and adapter size of 4. Distributed training ensured memory efficiency and faster convergence.

5.3. Data Augmentation

To mitigate the limitations of a relatively small training set, we employed translation-based data augmentation. Data augmentation is a well-established method in NLP to artificially expand datasets, and it is particularly effective in low-resource language scenarios where annotated corpora are scarce [81].

In our setup, Hungarian sentences were translated into English using the Google Translate API (https://cloud.google.com/translate (accessed on 30 May 2026)), creating additional training examples that are semantically equivalent but linguistically distinct from the originals. To prevent information leakage, the Hungarian corpus was first divided into fixed train, validation, and test subsets. Machine translation was then applied separately within each already-fixed subset, rather than to the complete corpus before splitting. This order is methodologically important because translating the complete corpus before splitting could allow a Hungarian sentence and its English translation to appear in different subsets. The resulting leakage-safe augmentation pipeline is shown in Figure 1.

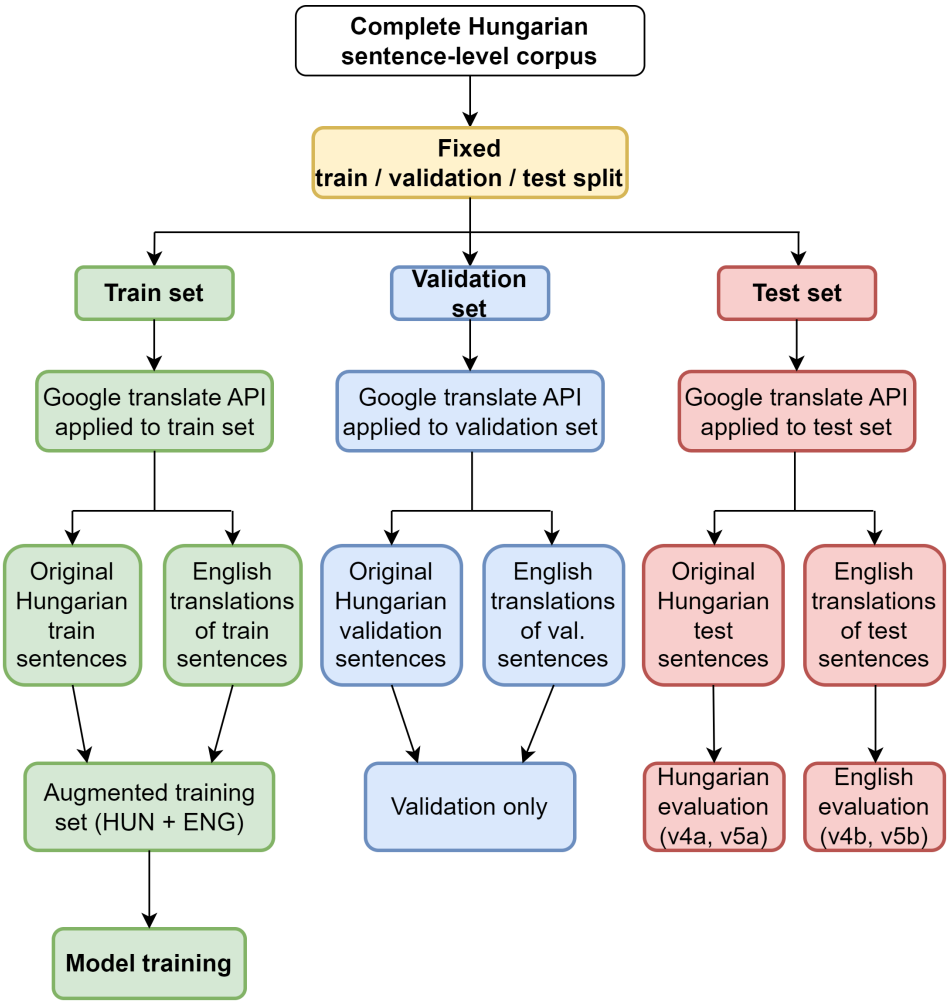

Figure 1. Translation-based augmentation pipeline. The Hungarian corpus was first split into fixed training, validation, and test subsets. Machine translation into English was then performed separately within each subset. Green, blue, and red boxes represent the training, validation, and test subsets, respectively.

This approach leverages the multilingual capabilities of models like XLM-RoBERTa, allowing them to learn from syntactic and lexical variation across languages [82,83]. Since English is over-represented in the pretraining corpora of most transformer models, the translated texts also align more closely with the models' internal representations, effectively enriching their understanding of clarity-related features [84].

This strategy offers three main benefits. First, it increases the volume of training data without requiring manual annotation. Second, it introduces diverse sentence structures and vocabulary that help the model generalize better. Third, it implicitly bridges the gap between underrepresented languages and better-resourced ones by aligning training data with the model's strongest linguistic priors.

However, translation-based augmentation is not without risks. Legal texts are particularly sensitive to semantic nuances, and machine translation may introduce artifacts that alter or obscure the original meaning. To mitigate this, we manually reviewed a subset of the translated texts to ensure that critical legal distinctions were preserved.

To further assess whether translation-based augmentation altered the linguistic feature space relevant to Plain Language classification, we conducted a lightweight diagnostic comparison between the original Hungarian sentences and their machine-translated English counterparts. We computed a set of surface-level proxy features for both languages, including sentence length, character count, comma density, long-word ratio, formal legal–administrative marker frequency, relative marker frequency, and a simple nominalization proxy. These features were not intended to provide a complete grammatical comparison between Hungarian and English sentences of the corpus. Rather, they were used to examine whether the contrast between Plain and Non-Plain sentences was preserved, weakened, or reconfigured after translation. For each feature, we compared class-wise means and Cohen's d effect sizes in the Hungarian and English versions of the corpus.

The metrics were calculated using simple and reproducible surface-level procedures. Sentence length was measured as the number of regex-based word tokens in each sentence, while character count was computed from the raw sentence string. Long-word ratio was defined as the proportion of tokens containing at least twelve characters. Comma density was calculated as the number of commas per 100 tokens. Formal legal–administrative marker frequency was based on a manually defined list of language-specific expressions that frequently occur in bureaucratic or legal–administrative prose. The Hungarian list included, among others, *amennyiben* [provided that/if], *esetén* [in case of/in the event of], *során* [during/in the course of], *tekintettel* [with regard to/taking into account], *vonatkozásában* [with respect to/regarding], *értelmében* [pursuant to/under], *történő* [being performed/carried out], and *minősül* [qualifies as/is deemed to be]. These items were selected because they often function as markers of formal administrative style, nominalized constructions, or light-verb-like bureaucratic phrasing. The English list included comparable administrative and legal expressions, such as *provided that*, *in the case of*, *during*, *with regard to*, *in respect of*, *pursuant to*, *concerning*, and *qualifies as*. Relative marker frequency was computed from occurrences of relative or subordinator-like markers, such as Hungarian *amely* [which], *amelynek* [whose/of which], *ahol* [where], and *amelyben* [in which], and English *which*, *that*, and *where*. The nominalization proxy was based on suffix-pattern matching for common nominal or gerund-like endings in each language.

For frequency-based features, raw counts were normalized per 100 tokens, calculated as follows:

$$\text{marker\_rate}_{100} = \frac{\text{marker\_count}}{\text{token\_count}} \times 100$$

where marker_rate_{100} denotes the normalized frequency of the relevant marker per 100 tokens, marker_count denotes the number of occurrences of that marker, and token_count denotes the total number of tokens in the sentence.

For each feature, we compared class-wise means and Cohen's d effect sizes in the Hungarian and English versions of the corpus. Cohen's d was calculated within each language as follows:

$$d = \frac{\bar{x}_{\text{Non-Plain}} - \bar{x}_{\text{Plain}}}{s_{\text{pooled}}}$$

where $\bar{x}_{\text{Non-Plain}}$ and $\bar{x}_{\text{Plain}}$ denote the mean feature values for the Non-Plain and Plain classes, respectively, and s_{pooled} denotes the pooled standard deviation of the two classes within the same language. These measures are therefore not intended as language-independent grammatical annotations, but as transparent diagnostic proxies for assessing whether class-separating surface cues are preserved or weakened after translation.

Incorporating translated data into the training pipeline ultimately improved the models' ability to detect stylistic and structural patterns characteristic of PL, even when applied to Hungarian content.

5.4. Contextual Baseline Reference

The main experimental comparison in the present study focuses on zero-shot prompting and fine-tuned transformer-based models. However, to avoid interpreting the zero-shot GPT-4o setup as the only non-fine-tuned point of reference, we also relate our findings to previously published lightweight machine learning baselines developed on the same corpus and sentence-level Plain Language classification task [13]. That earlier study evaluated term frequency–inverse document frequency (TF-IDF) features combined with a support vector machine (SVM), as well as fastText models using the same core corpus construction logic, namely the distinction between original administrative-legal sentences and their expert-rewritten Plain Language counterparts.

These previously published results are used here as contextual baselines rather than as strict head-to-head comparisons. This distinction is important because the earlier experiments were not re-run under the exact same train–test split and evaluation pipeline as the transformer models reported in the present study. Nevertheless, they provide a relevant lightweight supervised reference point, especially because the earlier SVM and fastText experiments included systematic hyperparameter optimization. The numerical comparison with these baselines is therefore presented in the Discussion after the current transformer-based results have been reported.

6. Results

This section presents the evaluation results of six different model configurations applied to the task of PL classification in Hungarian legal/official texts. The analysis begins with a zero-shot prompting baseline using the GPT-4o model, followed by fine-tuned model variants including huBERT, XLM-RoBERTa (base and large), GPT-4o-mini, and Gemini 1.0 Pro. We also assess the impact of data augmentation via Hungarian-to-English machine translation on model performance. The results are compared using standard classification metrics, with a focus on macro F1-scores, accuracy, and model robustness. Finally, we discuss observed trade-offs between model size, training strategy, and inference efficiency.

6.1. Zero-Shot Classification Results

As a baseline, we evaluated the GPT-4o model using a zero-shot prompting approach, in which the model was asked to classify legal sentences as either *Plain* or *Non-Plain* based solely on a natural language instruction. No task-specific fine-tuning or training examples were provided. To ensure deterministic outputs, the model was queried with `temperature=0`, which allowed for consistent and reproducible classification.

Table 3 summarizes the results. The model achieved an overall accuracy of 52%, which is only marginally better than random guessing for a balanced binary task. A breakdown by class reveals asymmetries: the model was somewhat better at identifying *Plain Language* texts, with a precision of 53% and a recall of 58%, resulting in an F1-score of 55%. By contrast, performance on *Non-Plain* sentences was weaker (F1 = 48%), with both precision and recall below 52%.

Table 3. Evaluation metrics for zero-shot prompting (GPT-4o).

Category	Precision	Recall	F1-Score
Plain	53	58	55
Non-Plain	51	46	48
Accuracy		52	
Macro avg	52	52	52
Weighted avg	52	52	52

This mild bias toward the Plain category may stem from the phrasing of the prompt or from the model's general preference for more neutral or accessible formulations when uncertainty arises. Similar tendencies have been observed in prior studies, where instruction-tuned language models without domain supervision tend to produce or favor simpler outputs in ambiguous contexts [85,86].

While zero-shot prompting offers practical advantages, such as eliminating the need for annotated training data, its performance here demonstrates the limitations of generic language models in domain-specific classification tasks. The low F1-scores and near-chance accuracy indicate that GPT-4o lacks a sufficiently nuanced representation of legal linguistic complexity when operating without additional guidance.

These results underscore the necessity of supervised fine-tuning for effective PL detection in legal texts. In the following sections, we explore how model performance improves when trained on aligned Hungarian corpora containing Plain and Non-Plain versions of legal–administrative content.

6.2. Fine-Tuning Results

To evaluate the effect of supervised adaptation, we fine-tuned a series of transformer-based models using Hungarian legal texts annotated for PL. Table 4 provides an overview of the model configurations, training regimes, and evaluation languages. Each setup is assigned a version identifier (v1–v7) to facilitate consistent cross-reference in the Sections 6 and 7.

Table 4. Overview of fine-tuned model configurations. Data augmentation via machine translation was applied in versions v4–v5. The train/validation/test split was fixed before translation. The "Evaluated on" column refers to the language of the input used during testing: HUN indicates original Hungarian sentences, ENG indicates their machine-translated English counterparts.

Version	Model	Params (M)	Data Size	Train Lang.	Evaluated on
v1	huBERT	111	11k	HUN	HUN
v2	XLM-RoBERTa-base	278	11k	HUN	HUN
v3	XLM-RoBERTa-large	560	11k	HUN	HUN
v4a	XLM-RoBERTa-base	278	22k (11k + 11k)	HUN + ENG	HUN
v4b	XLM-RoBERTa-base	278	22k (11k + 11k)	HUN + ENG	ENG
v5a	XLM-RoBERTa-large	560	22k (11k + 11k)	HUN + ENG	HUN
v5b	XLM-RoBERTa-large	560	22k (11k + 11k)	HUN + ENG	ENG
v6	GPT-4o-mini	not disclosed	11k	HUN	HUN
v7	Gemini 1.0 Pro	not disclosed	11k	HUN	HUN

Table 4 describes the main parameters of the carried out experiments. In v4 and v5, data augmentation was applied after the train/validation/test split had been fixed. The validation and test subsets were translated separately and were not added to the training data. The Hungarian evaluation settings, v4a and v5a, used the original Hungarian held-out test sentences, whereas the English evaluation settings, v4b and v5b, used the machine-translated English counterparts of the same fixed test split. This setup enabled models to benefit from multilingual representations, particularly in XLM-RoBERTa, which was pretrained on over 100 languages, without data leakage. Evaluating on machine-translated English sentences offers insight into the model's cross-lingual robustness, especially given that English is significantly over-represented in the pretraining corpus of multilingual transformers like XLM-RoBERTa.

Before fine-tuning, we assessed whether the 512-token input limit imposed by huBERT and XLM-RoBERTa would lead to data loss. Empirical analysis showed this was negligible: only 15 sentences (0.13%) in huBERT and 23 sentences (0.21%) in XLM-RoBERTa exceeded the limit out of 10,928 examples. Average input lengths were 47.63 and 57.82 tokens, respectively, confirming that truncation had minimal effect on semantic coverage.

Table 5 summarizes evaluation scores for each model version. The best-performing setup, v5a (XLM-RoBERTa-large with multilingual training, evaluated on Hungarian), achieved the highest macro-average F1-score (0.79). This confirms that large multilingual transformers benefit from cross-lingual fine-tuning, even when evaluation remains monolingual.

Table 5. Evaluation metrics for fine-tuned models. Best results per metric are highlighted in grey.

Metric/Model	v1	v2	v3	v4a	v4b	v5a	v5b	v6	v7
Evaluated on	**HUN**	**HUN**	**HUN**	**HUN**	**ENG**	**HUN**	**ENG**	**HUN**	**HUN**
Plain P	0.71	0.68	0.74	0.82	0.70	**0.82**	0.68	0.71	0.63
Plain R	**0.79**	0.67	0.65	0.62	0.50	0.74	0.60	0.66	0.52
Plain F1	0.75	0.67	0.70	0.70	0.58	**0.78**	0.64	0.68	0.57
Non-Plain P	0.76	0.69	0.71	0.71	0.63	**0.77**	0.66	0.69	0.60
Non-Plain R	0.67	0.71	0.79	**0.88**	0.80	0.85	0.73	0.73	0.70
Non-Plain F1	0.71	0.70	0.74	0.78	0.70	**0.81**	0.69	0.71	0.65
Macro P	0.74	0.69	0.73	0.77	0.66	**0.80**	0.67	0.70	0.62
Macro R	0.73	0.69	0.72	0.75	0.65	**0.79**	0.67	0.70	0.61
Macro F1	0.73	0.69	0.72	0.74	0.64	**0.79**	0.67	0.70	0.61

In contrast, evaluations performed on the English translations (v4b, v5b) yielded slightly lower macro-average F1-scores (0.64 and 0.67, respectively). These results suggest that while cross-lingual fine-tuning can improve general robustness, model performance remains sensitive to the language of evaluation. Machine-translated English inputs may introduce subtle inconsistencies or stylistic artifacts that negatively impact classification accuracy. Furthermore, although XLM-RoBERTa was pretrained on English data extensively, the original sentence structure and legal phrasing are rooted in Hungarian, which may reduce semantic fidelity when evaluated in English. This reinforces the importance of training and evaluating PL classifiers primarily in the language of deployment.

Compared to the zero-shot GPT-4o baseline (macro F1 = 0.52), all fine-tuned models demonstrate substantial performance gains, underlining the importance of domain-specific adaptation.

Among all configurations, the v5a model (i.e., XLM-RoBERTa-large trained on both Hungarian texts and their machine-translated English counterparts) achieves the strongest overall performance. It obtains a macro-average F1-score of 0.79, with balanced results

across both classes. Notably, it also performs best at detecting Non-Plain sentences, achieving the highest F1-score (0.81) and precision (0.77) in this category.

The improvement from v3 to v5a (both using XLM-RoBERTa-large, but with and without data augmentation) illustrates the effectiveness of cross-lingual fine-tuning, while v3 reaches a macro F1-score of 0.72, v5a improves this to 0.79. Similar gains are observed when comparing v2 and v4a (XLM-RoBERTa-base with and without data augmentation), where the macro F1-score increases from 0.69 to 0.74. This demonstrates that data augmentation not only boosts overall accuracy but is particularly effective for identifying complex, inaccessible sentence structures.

The GPT-4o-mini model (v6), trained and evaluated entirely on Hungarian data, delivers competitive results (macro F1 = 0.70). This suggests that instruction-tuned LLMs can generalize reasonably well to PL classification, even in a monolingual, domain-specific setting, without access to multilingual augmentation.

The Gemini 1.0 Pro model (v7), also trained on Hungarian-only data, performs moderately (macro F1 = 0.61). It shows better recall for Non-Plain cases (0.70), but underperforms on Plain sentences (F1 = 0.57), indicating a conservative bias towards complexity. Compared to open-weight alternatives with data augmentation, its generalization appears more limited in this legal domain.

To summarize, fine-tuning large multilingual transformer models with data augmentation yields the best results. However, lightweight monolingual models like huBERT still offer competitive performance (macro F1 = 0.73), and minimal token truncation was observed in less than 0.2% of the inputs, indicating robustness across all configurations.

6.3. Model Size, Efficiency, and Deployment Trade-Offs

Figure 2 visualizes the trade-off between model size (in millions of parameters), classification performance (macro-average F1 score), and inference time. Each bubble represents a fine-tuned model. This combined visualization supports practical trade-off decisions between accuracy, latency, and model scalability, especially in domains with strict resource constraints.

The results highlight key considerations in the ongoing debate between Small and Large Language Models [87]. On the one hand, larger models such as v5a(XLM-RoBERTa-large trained on Hungarian and translated English texts) achieve the best overall performance (macro F1 = 0.79), while still maintaining acceptable inference times. However, the high computational demands of these models pose challenges for deployment, particularly in resource-constrained or latency-sensitive environments.

Smaller models, including v1 (huBERT) and v4a (XLM-RoBERTa-base with data augmentation), demonstrate competitive performance (macro F1 = 0.73 and 0.74, respectively), while offering clear advantages in cost-efficiency, and deployability. These characteristics are particularly relevant in legal and institutional settings where data privacy, auditability, and infrastructure constraints are non-negotiable. Notably, v1 (huBERT) achieves strong performance despite having the fewest parameters and the lowest inference latency, illustrating that small models can remain viable when domain-specific training is applied.

The proprietary GPT-4o-mini (v6) model, while performing reasonably well (macro F1 = 0.70), is accessible only through OpenAI's API, limiting its applicability due to permanent usage costs, lack of transparency, and restrictions on model customization. Similar considerations apply to the Gemini 1.0 Pro model (v7), which also operates as a closed-source system. Despite its smaller training scope and absence of data augmentation, Gemini achieves a macro F1 of 0.61, reflecting moderate generalization capabilities within the Hungarian official domain.

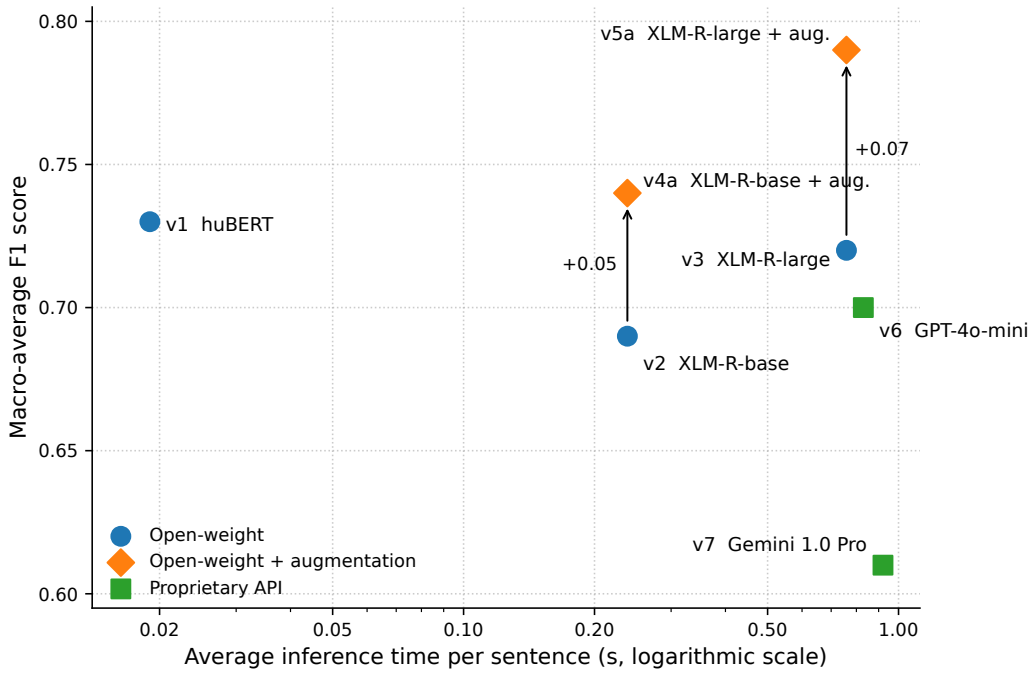

Figure 2. Trade-off between average inference time and macro-average F1 score. The x-axis uses a logarithmic scale to improve the visual separation of low-latency models. Arrows indicate the performance gains obtained through Hungarian–English translation-based data augmentation.

Inference times (as detailed in Table 6) for each model were measured during prediction on the held-out test set used for evaluation. Average latency values were computed over all test examples, reflecting realistic end-to-end response times under practical deployment conditions. These values capture the combined effect of model architecture, inference environment, and backend optimizations. Inference was performed using a single NVIDIA A100 GPU unit (NVIDIA Corporation, Santa Clara, CA, USA) with batch size = 1 to approximate real-time API usage scenarios. For proprietary models (v6 and v7), which are only accessible via API, latency measurements also include request transmission time and server-side processing delays. This makes them less predictable and potentially less suitable for latency-critical environments.

Table 6. Average inference time per input example for each fine-tuned model.

Model Version	Model	Average Inference Time (s)
v1	huBERT	0.019
v2	XLM-RoBERTa-base	0.238
v3	XLM-RoBERTa-large	0.759
v4	XLM-RoBERTa-base (aug)	0.238
v5	XLM-RoBERTa-large (aug)	0.759
v6	GPT-4o-mini	0.830
v7	Gemini 1.0 Pro	0.921

These findings underline the recurring trade-offs between model accuracy, latency, and deployability; while instruction-tuned proprietary LLMs can offer strong performance, they introduce challenges in terms of cost, access, and long-term sustainability. Conversely, open-weight models hosted on platforms such as Hugging Face enable fine-tuning on task-specific data and self-hosted deployment at the same time, providing greater control over data security, compliance, and customization.

In summary, results suggest that mid-sized multilingual models with augmentation (e.g., v4a, v5a) offer a promising compromise, meriting further attention in practical deployments. They combine strong performance with manageable resource requirements, making them especially well-suited for high-stakes domains like legal text classification, where accuracy must be balanced with interpretability, compliance, and technical viability.

7. Discussion

This section synthesizes the empirical results presented above to derive theoretical and practical insights. We examine the relative effectiveness of different model architectures, the role of data augmentation, and the challenges associated with PL classification in low-resource legal domains. Limitations and implications for future research and deployment are also discussed.

7.1. Interpretation of Model Performance

The evaluation results presented in Table 5 indicate that multilingual transformer models, particularly XLM-RoBERTa-large trained with augmented Hungarian–English data (v5a), outperform other configurations in classifying legal and administrative texts according to PL standards. These findings support prior evidence that multilingual pretraining can enhance downstream task performance in low-resource languages through improved representation learning and domain transfer [83].

The v5a model achieved a macro-average F1-score of 0.79, indicating a strong ability to distinguish between Plain and Non-Plain sentence constructions. Its consistent performance across precision and recall, especially for Non-Plain sentences, suggests reliable identification of linguistically complex inputs. This likely reflects the model's exposure to syntactically and lexically diverse inputs, enabled by machine-translated English augmentations, which may have promoted more generalized representations of sentence clarity.

By contrast, the zero-shot GPT-4o model achieved only near-random performance (macro F1 = 0.52), underscoring the limitations of relying solely on general-purpose instruction-tuned pretraining without task-specific adaptation. Despite its extensive multilingual training, GPT-4o lacked supervised exposure to Hungarian legal texts, which appears essential for accurate PL classification.

Notably, the huBERT model (v1), a monolingual, compact model pretrained on Hungarian corpora, achieved a macro F1-score of 0.73. This performance demonstrates the value of domain-specific pretraining, showing that even relatively small models can approach the effectiveness of larger architectures when fine-tuned on curated in-domain data. Given its low inference latency and open-weight availability, huBERT represents a viable solution for deployment in resource-constrained or regulated environments, such as legal or institutional settings.

7.2. Comparison with Previously Published Lightweight Baselines

The present results can also be interpreted in relation to previously published lightweight baselines on the same corpus and sentence-level Plain Language classification task [13]. This comparison is useful because the zero-shot GPT-4o result alone would provide only a weak baseline for assessing the value of supervised transformer fine-tuning. The earlier study evaluated two non-transformer approaches: TF-IDF + SVM and fastText. Although these models were not re-run under the exact same split as the current experiments, they were trained and evaluated on the same corpus and therefore provide a meaningful contextual reference point.

In the earlier TF-IDF + SVM experiment, the regularization parameter C, the γ parameter, and the kernel type were systematically optimized. The tested values were

$C \in 0.1, 1, 10, 100$, $\gamma \in 1, 0.1, 0.01, 0.001$, and the rbf, poly, sigmoid, and linear kernels. Model performance was evaluated using 10-fold cross-validation. The best SVM configurations achieved an average two-class F1-score of approximately 0.68.

The same earlier study also evaluated fastText models with both pretrained Hungarian vectors and vectors trained on the task-specific corpus. The fastText experiments included hyperparameter testing for the learning rate, word n-gram size, minimum token count, embedding dimensionality, and number of epochs. The best configuration used a learning rate of 0.1, unigram features, a minimum token count of 20, and 50-dimensional embeddings. However, the best fastText models reached only approximately 0.62 F1-score, underperforming the optimized SVM baseline.

Taken together, these earlier results show that the task is learnable with lightweight supervised methods, but they also suggest a lower performance ceiling for traditional approaches. The comparison is summarized in Table 7. The present best-performing transformer model, XLM-RoBERTa-large trained with Hungarian–English augmentation, achieved a macro-average F1-score of 0.79. This indicates that the performance gain of supervised transformer fine-tuning is not limited to the comparison with zero-shot GPT-4o; it also exceeds previously published lightweight supervised baselines that had already undergone hyperparameter optimization.

Table 7. Contextual comparison with previously established lightweight baselines.

Model/Setup	Optimization	Evaluation Setting	Best Reported Result
TF-IDF + SVM	C, γ, and kernel type	10-fold cross-validation in the earlier NAV-based baseline study	Approx. 0.68 average two-class F1
fastText	Learning rate, word n-grams, minimum token count, embedding dimension, and epochs	train–test evaluation with repeated runs across epochs in the earlier baseline study	Approx. 0.62 F1
GPT-4o zero-shot	No task-specific training	Held-out test set in the present study	0.52 macro F1
XLM-RoBERTa-large + augmentation	Supervised transformer fine-tuning with Hungarian–English augmentation	Held-out Hungarian test set in the present study	0.79 macro F1

7.3. Impact of Data Augmentation

Translation-based data augmentation significantly enhanced model performance across multiple configurations. Both XLM-RoBERTa-base (v2 vs. v4a) and XLM-RoBERTa-large (v3 vs. v5a) exhibited measurable improvements when trained on augmented corpora, with macro F1-scores increasing by 0.05 and 0.07, respectively. These gains suggest that exposing models to syntactic variability, even via machine-translated text, improves their capacity to abstract the linguistic features that define PL.

These gains should be interpreted in light of the leakage-safe augmentation design described in Section 5.3. Since the split was fixed before translation, the observed improvement cannot be attributed to the model seeing translated variants of validation or test sentences during training. Instead, the improvement reflects the additional cross-lingual training signal provided by the English translations of the training split only.

To better understand the linguistic consequences of this augmentation strategy, we examined whether selected surface-level PL-related proxy features behaved similarly in the Hungarian originals and in their English translations. The results are summarized in

Table 8. The diagnostic comparison suggests that translation did not simply shorten or simplify the sentences in a uniform way. On average, the English translations contained more tokens than the Hungarian originals, with an English-to-Hungarian token ratio of 1.36 for Plain sentences and 1.34 for Non-Plain sentences. However, the class-separating effect of certain Hungarian legal–administrative markers became substantially weaker after translation. In Hungarian, formal markers per 100 tokens showed a clear difference between Plain and Non-Plain sentences: 1.36 in Plain sentences and 2.86 in Non-Plain sentences, with a Cohen's d of 0.47. In English, the corresponding values were 2.29 and 2.64, with a much smaller Cohen's d of 0.10. This indicates that the translation process may smooth out some of the surface-level cues that distinguish Non-Plain Hungarian legal–administrative style from Plain Language.

This shift is linguistically plausible. Hungarian is an agglutinative language in which case suffixes, possessive constructions, verbal prefixes, nominalizations, and flexible word order can encode dense legal–administrative relations compactly. In English, these relations are often rendered through prepositional phrases, more fixed word order, and explicit subordination. As a result, machine translation may preserve the general semantic content while redistributing the formal and structural cues that the classifier can use. The increase in relative markers in English further supports this interpretation: some compact Hungarian constructions appear to be mapped into more explicit analytic structures, such as relative or subordinator-based formulations. These findings do not invalidate translation-based augmentation, but they show that its benefits should be interpreted as cross-lingual regularization rather than as a neutral expansion of the Hungarian feature space.

Table 8. Diagnostic comparison of selected surface-level Plain Language proxy features before and after machine translation. Values are computed on the aligned training split. Cohen's d expresses the Plain vs. Non-Plain effect size within each language.

Feature	HUN Plain	HUN Non-Plain	HUN d	ENG Plain	ENG Non-Plain	ENG d
Sentence length, tokens	26.56	29.89	0.13	35.75	39.84	0.12
Formal markers per 100 tokens	1.36	2.86	0.47	2.29	2.64	0.10
Formal marker count	0.42	0.90	0.42	0.81	1.04	0.17
Relative marker count	0.08	0.11	0.08	0.32	0.42	0.13
Nominalization proxy count	1.34	1.65	0.13	2.74	3.20	0.13

However, the performance drop in evaluations conducted on English-translated test data (v4b, v5b) highlights an important caveat: data augmentation is most effective when limited to the training phase. Evaluation in the language of deployment (in this case, Hungarian) remains crucial. Translation artifacts, such as altered idiomaticity or inconsistent legal terminology, may distort model predictions and reduce ecological validity. Consequently; while augmentation supports model generalization, deployment pipelines should preserve original language inputs to ensure faithful classification.

In practice, this means that multilingual models like XLM-RoBERTa can benefit substantially from English augmentations during training, but inference and validation should rely on the source-language corpus for accuracy and domain fidelity.

7.4. Model Bias and Label Imbalance

Across fine-tuned configurations, a recurring pattern emerged: models generally performed better at detecting Non-Plain constructions than Plain ones. This asymmetry suggests the presence of underlying biases in either the data or the model representations.

One plausible explanation is that Non-Plain sentences tend to exhibit more salient structural markers (nested clauses, nominalizations, or passive constructions) that are

readily captured by transformer-based encoders. These features may act as strong signals for "complexity," facilitating classification.

Another contributing factor could be annotation bias. If human annotators were more confident in identifying complex sentences (e.g., based on surface difficulty or bureaucratic tone) than in positively labeling clear, accessible formulations, the resulting data distribution might have skewed the learning process. This would lead models to develop a heightened sensitivity to markers of Non-Plainness, while under-representing the diverse features of effective PL.

Addressing this imbalance will likely require both data-level and model-level interventions. On the data side, targeted augmentation of underrepresented Plain examples, especially those with varied syntax and professional phrasing, could help balance the distribution. On the model side, techniques from explainable AI (XAI), such as gradient-based saliency maps or SHAP analyses, could be employed to better understand and potentially correct internal biases by revealing the features most influential in classification decisions.

Finally, future work may also explore cost-sensitive training or calibrated loss functions to explicitly account for asymmetries in label distribution and decision importance, especially in real-world deployments where misclassifying a Plain sentence might have lower risk than overlooking a complex one.

7.5. Interpretability via SHAP Analysis

To better understand the decision-making heuristics of the best-performing model (v5–XLM-RoBERTa-large with augmented data), we conducted local interpretability analysis using SHAP (SHapley Additive exPlanations) [16]. SHAP is a model-agnostic framework based on cooperative game theory that attributes to each input token a contribution value toward the model's prediction. This enables quantifying and visualizing the influence of individual tokens, thereby enhancing transparency and linguistic interpretability.

In addition to the local SHAP explanations, we also performed an aggregated word-level SHAP analysis on the full Hungarian held-out test set. This additional analysis was conducted for the two open-weight models most relevant to deployment and comparison:

- The monolingual huBERT-based model (v1).
- The best-performing XLM-RoBERTa-large model trained with Hungarian–English augmentation (v5).

Sentences were grouped according to the models' own predicted class labels. For each model and predicted class, we aggregated only positive SHAP contributions toward the predicted class. To avoid tokenizer-specific artifacts, contributions were aggregated at the surface-word level rather than at the subword-token level. Only words with at least five positive SHAP occurrences were retained, and both models' predictions were evaluated solely on the Hungarian test set.

7.5.1. True Positive Example: Nominalized and Bureaucratic Constructions

Here, the term true positive is used with *Non-Plain* as the positive class. To illustrate linguistic complexity in Hungarian legal–administrative style, consider the following example:

> *Kivételt képez az újrahasználható raklap engedélyezett bérleti rendszer üzemeltetője által, bérleti rendszer keretén belüli használat céljából, rendelkezésre bocsátása.*
>
> *[Literal translation: An exception is constituted by the making available of the reusable pallet by the operator of the authorized leasing system, for the purpose of use within the framework of the leasing system].*

The sentence exhibits features typical of legal–bureaucratic registers. Here, extended nominal constructions, such as *keretén belüli használat* [use within the framework], abstract

nominal expressions, such as *rendelkezésre bocsátása* [making available], and oblique syntactic structures that hinder comprehension.

Figure 3 presents the local SHAP explanation for the model's prediction. The sentence was correctly classified as *Non-Plain*, with a predicted probability of 0.9963 for the *Non-Plain* class. This probability was derived by applying a softmax function to the model's logits. The explanation uses the verified class-index mapping used in the SHAP analysis, where `output_0` corresponds to the *Non-Plain* class. The figure uses a hybrid visualization. The upper panel preserves the full token sequence and colors each token or subword unit according to its local SHAP contribution. The lower panel ranks the strongest positive and negative local SHAP contributors. Red elements increase the model's confidence in the *Non-Plain* prediction, while blue elements push the prediction in the opposite direction.

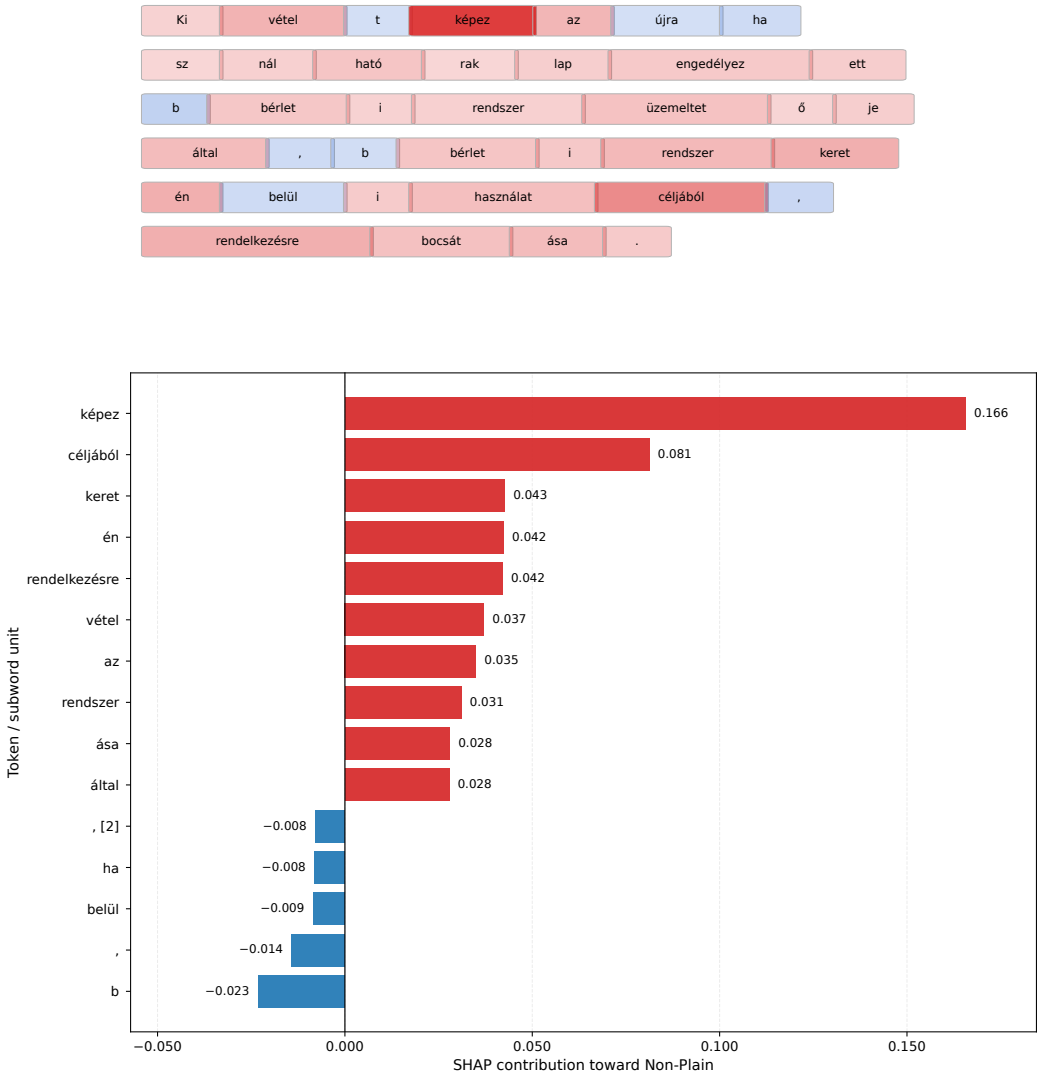

Figure 3. Hybrid local SHAP explanation for a correctly classified *Non-Plain* sentence. The (**upper**) panel shows the full token sequence, while the (**lower**) panel ranks the strongest local SHAP contributors toward the *Non-Plain* class. Red elements contribute positively toward the *Non-Plain* class, while blue elements contribute negatively with respect to that class. The notation "[2]" is indicates that the corresponding comma token is the second occurrence of the same punctuation mark in the original sentence. The fragments shown in y-axis in the visualization are Hungarian token and subword units generated by the model tokenizer from the original input sentence; individual subword fragments may not have an independent lexical meaning.

The strongest positive contributors include token and subword units belonging to expressions such as *kivételt képez* [constitutes an exception], *bérleti rendszer keretén belüli használat céljából* [for the purpose of use within the framework of the leasing system], and *rendelkezésre bocsátása* [making available]. In the ranked panel, particularly strong contributions are associated with units such as *képez*, *céljából*, *keret*, *rendelkezésre*, *rendszer*, *ása*, and *által*. These items should not be interpreted as isolated lexical triggers. Because the local explanation is produced at token or subword level, some units become linguistically meaningful only when they are read as parts of complete surface words or larger constructions. Their relevance is therefore best understood in the context of nominalized, periphrastic, and bureaucratic phrasing.

Such expressions typify bureaucratic language patterns that reduce accessibility. First, nominalizations like *rendelkezésre bocsátása* [making available] introduce abstraction that could often be avoided with verb-based forms, such as *elérhetővé teszi* [makes it accessible]. Second, embedded syntactic structures, such as *keretén belüli használat* [use within the framework], increase cognitive load. Third, institutional and purpose-marking expressions such as *céljából* [for the purpose of] distance the text from everyday usage.

Especially notable is the light verb construction [88] *kivételt képez* [constitutes an exception], which makes the sentence more indirect and could be more plainly expressed, for example as *nem tartozik bele* [is not included]. Such periphrastic patterns are prevalent in legal drafting and contribute significantly to perceived textual complexity.

These linguistic features align with the patterns the model identified as characteristic of the *Non-Plain* class. The hybrid visualization makes this relationship clear because it preserves the token sequence while also presenting the strongest SHAP contributors in a readable ranked form.

7.5.2. Correctly Classified Plain Example: Domain-Specific Terminology Without Non-Plain Classification

Figure 4 presents a correctly classified *Plain* sentence containing several domain-specific legal–administrative terms. The example illustrates that the presence of technical or institutional vocabulary does not necessarily result in a *Non-Plain* prediction. Instead, the model appears to evaluate such lexical cues in combination with the broader syntactic and contextual structure of the sentence.

> *A tevékenységét év közben kezdő kkt., bt., egyéni cég a cégbírósági nyilvántartásba vételi kérelemmel egyidejűleg nyilatkozik, és az adatok az úgynevezett egyablakos rendszeren keresztül érkeznek meg a NAV-hoz.*
>
> *[Translation: A general partnership (kkt.), limited partnership (bt.), or sole proprietorship starting its activities during the year makes a declaration along with its company registration application, and the data are transmitted to the tax authority (NAV) through the so-called one-stop system].*

The sentence contains formal administrative vocabulary, including company-law abbreviations, *cégbírósági nyilvántartásba vételi kérelem* [company registration application], *egyablakos rendszer* [one-stop system], and *NAV* [National Tax and Customs Administration]. However, these terms appear in a relatively linear and interpretable sentence structure. The model classified the sentence as *Plain*, with a predicted probability of 0.9898 for the *Plain* class.

For comparability with the previous example, Figure 4 explains local SHAP contributions toward the *Non-Plain* class rather than toward the predicted *Plain* class. Therefore, red elements indicate token or subword units that would increase the model's confidence in a *Non-Plain* classification, while blue elements indicate units that push against the *Non-Plain* class. This makes it possible to examine why the sentence contains some local evidence

associated with administrative complexity, but still does not cross the model's decision boundary for *Non-Plain* language.

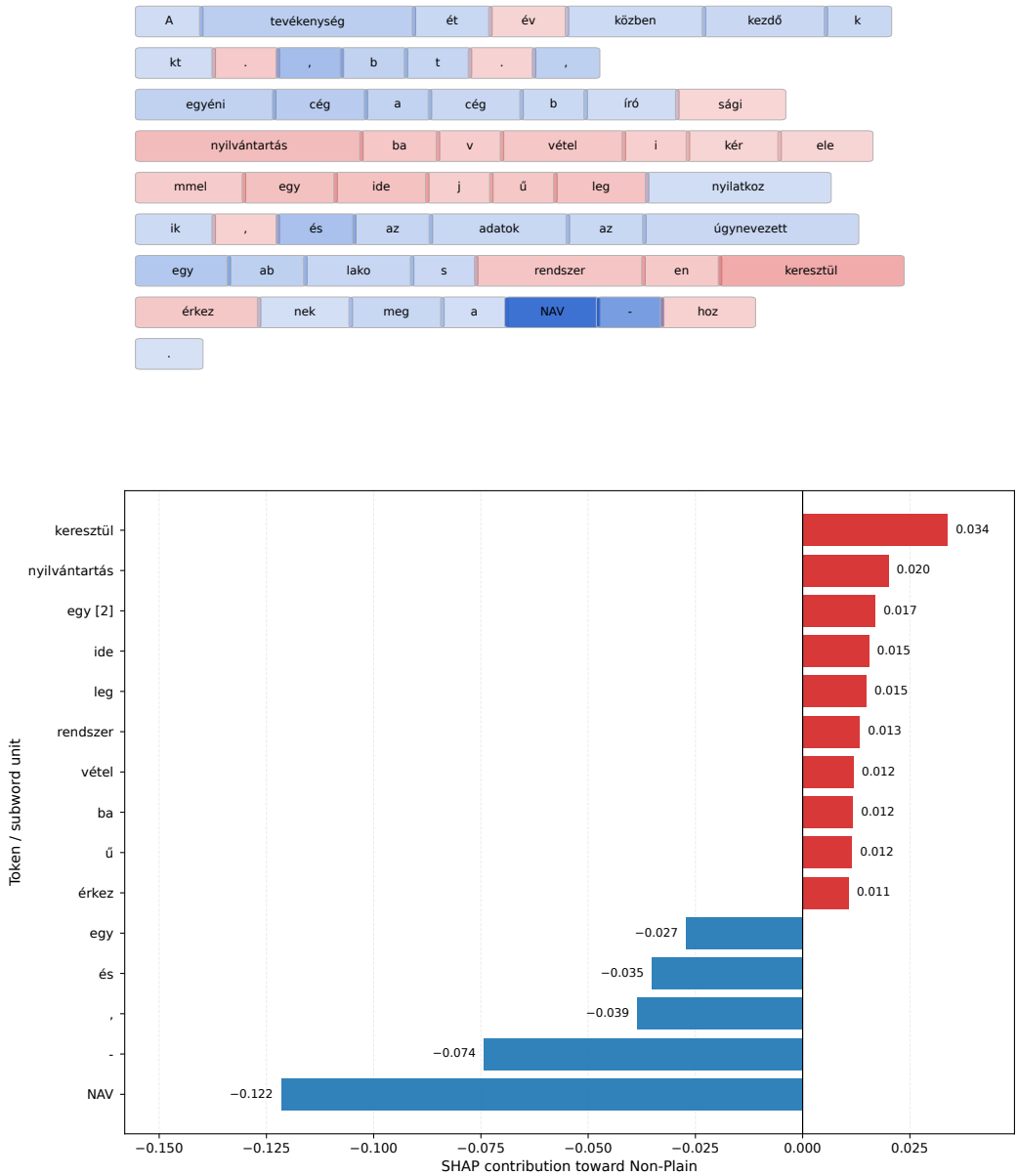

Figure 4. Hybrid local SHAP explanation for a correctly classified *Plain* sentence containing domain-specific legal–administrative terminology. The visualization explains contributions toward the *Non-Plain* class. Red elements contribute positively toward the *Non-Plain* class, while blue elements contribute negatively with respect to that class. The notation "[2]" is indicates that the corresponding comma token is the second occurrence of the same punctuation mark in the original sentence. The fragments shown in y-axis in the visualization are Hungarian token and subword units generated by the model tokenizer from the original input sentence; individual subword fragments may not have an independent lexical meaning.

The SHAP values show a mixed pattern. Some units, such as *keresztül* [through], *nyilvántartás* [registration/register], *vétel* [taking/registration-related nominal element], and *rendszer* [system], contribute positively toward the *Non-Plain* class. These items plausibly reflect formal administrative register and institutional procedure. At the same time, other elements counteract a *Non-Plain* interpretation. Most importantly, *NAV* appears as a strong negative contributor with respect to the *Non-Plain* class in this local explanation. This indicates that the acronym did not operate here as a simple lexical trigger for *Non-Plain* classification.

This contrastive example refines the interpretation of model behavior. The model does not appear to rely exclusively on isolated domain-specific terms. Instead, it combines lexical register cues with broader contextual and structural signals. The presence of institutional vocabulary alone is therefore not sufficient to produce a *Non-Plain* classification. This is important for legal–administrative Plain Language detection, because genuinely plain public-facing texts may still need to contain precise institutional and legal terminology.

At the same time, the example also shows why model auditing remains necessary. Several administrative terms still provide positive local evidence toward the *Non-Plain* class, even though the final prediction is *Plain*. This suggests that additional training data containing clear sentences with domain-specific terminology would be valuable. Such data could help the model further distinguish between unavoidable technical vocabulary and genuinely inaccessible legal–bureaucratic constructions.

Moreover, it is important to note that the dataset used for fine-tuning was constructed from sentence-level edits aimed at improving clarity; while the primary objective of these edits was to promote compliance with PL principles, the underlying editorial workflow may have introduced additional linguistic modifications not strictly tied to PL criteria. As a result, certain training examples may reflect broader stylistic or structural changes, inadvertently introducing noise into the label distribution. Recognizing this helps contextualize individual model decisions and guides future corpus design toward more annotation-consistent data curation.

7.5.3. Aggregated Word-Level SHAP Analysis

The two local SHAP examples above illustrate how individual model decisions can be explained in linguistically meaningful terms. However, local explanations alone do not show whether the same patterns are used systematically across the test set. To address this limitation, we complemented the local explanations with an aggregated word-level SHAP analysis on the full Hungarian held-out test set. The analysis was conducted for the two most relevant open-weight models: v1, the huBERT-based model, and v5, the XLM-RoBERTa-large model trained with Hungarian–English augmentation.

The aggregation was performed at the surface-word level. We did not aggregate the models' internal subword tokens directly after tokenization. Instead, SHAP explanations were generated with a text masker that segmented the original input string into surface-word units using a regular-expression-based segmentation pattern. These surface-word units served as the maskable explanation units. During SHAP estimation, complete surface-word segments were masked in the original input string, producing surface-word-masked sentence variants. Each masked variant was then passed to the model's own tokenizer. Consequently, the transformer model still processed the input through its model-specific internal tokenization scheme, but the SHAP values were assigned back to the original surface-word units.

This distinction is important because huBERT and XLM-RoBERTa use different internal tokenizers. A direct comparison at the subword-token level would therefore partly reflect tokenizer-specific artifacts rather than linguistically interpretable units. The word-level masking strategy makes the two models more comparable. Here, surface words define the explanation units, while model-specific subword tokenization remains part of the normal forward pass. In this way, the procedure preserves the prediction behavior of each model while producing SHAP values that can be interpreted and aggregated at the word level.

Figure 5 summarizes this architecture. The figure separates the explanation level from the model-internal level. At the explanation level, surface words are identified and used as SHAP masking units. At the model-internal level, the surface-word-masked variants are tokenized and processed by each model according to its own tokenizer. The resulting

changes in prediction probabilities are then attributed back to the original surface-word units, which allows word-level SHAP contributions to be accumulated across the test set.

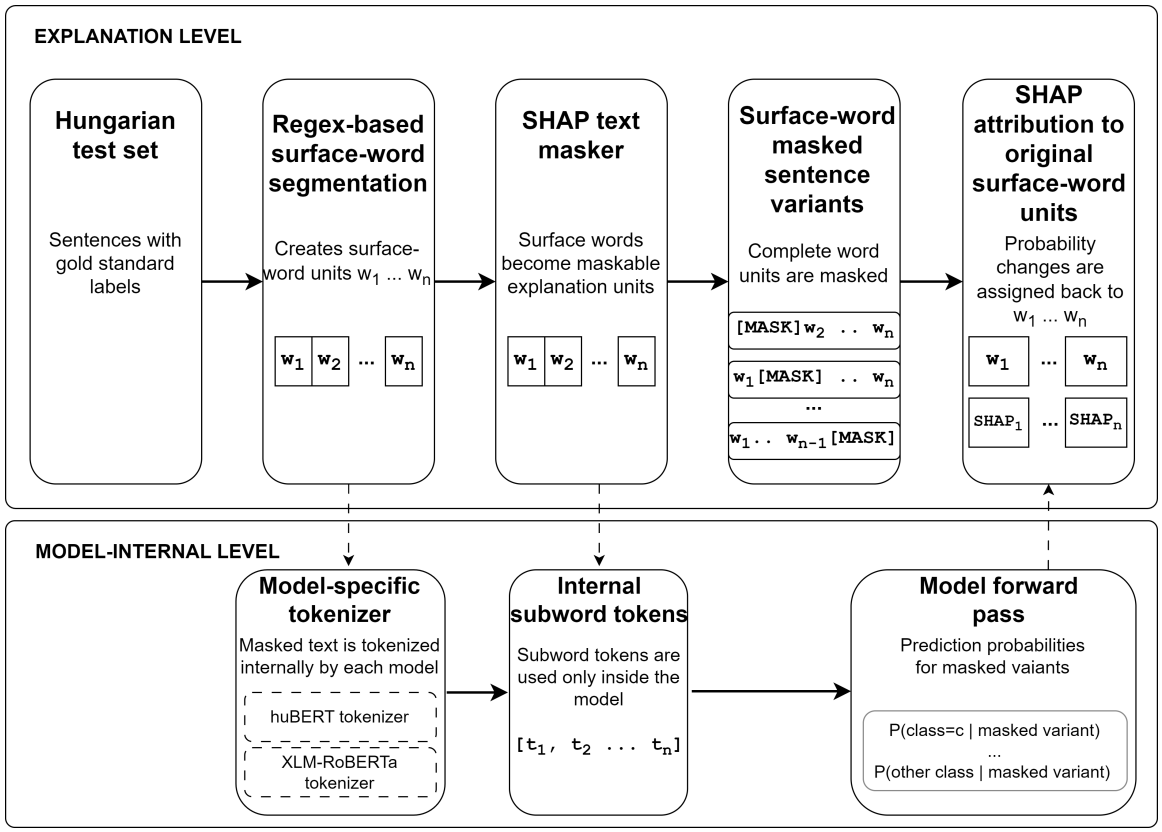

Figure 5. Architecture of the surface-word-level SHAP aggregation procedure. Input sentences are first segmented into surface-word units, which serve as the maskable explanation units. Surface-word-masked sentence variants are then passed to each model's own tokenizer and processed internally as model-specific subword sequences. SHAP attributions are computed from the resulting prediction changes and assigned back to the original surface-word units. These word-level contributions are then aggregated across the Hungarian held-out test set to compute mean positive SHAP scores for the predicted class.

For each sentence, we first identified the model's predicted class. We then retained only those SHAP values that contributed positively toward that predicted class. These positive word-level contributions were accumulated across all sentences assigned to the same predicted class. For each word, we calculated the mean positive SHAP contribution by dividing the sum of its positive SHAP values by the number of positive occurrences. We also recorded the total number of positive occurrences and the number of distinct sentences in which the word contributed positively. To reduce the influence of unstable rare items, only words with at least five positive SHAP occurrences were included in the final ranking.

The aggregated results (see Table 9) show a clear and linguistically interpretable pattern for Non-Plain predictions. In both models, several of the strongest contributors toward the Non-Plain class are formal legal–administrative expressions, including *tekintettel* [with regard to], *amennyiben* [insofar as/if], *során* [during], *történő* [performed/carried out], *vonatkozásában* [with respect to], *esetében* [in the case of], and *értelmében* [pursuant to/under]. These items are not necessarily incomprehensible in isolation, but they frequently occur in dense administrative constructions, nominalized phrases, and formal legal formulae. Their high aggregated SHAP values suggest that the models learned indicators associated with bureaucratic and legally formal sentence construction.

Table 9. Aggregated top-20 word-level SHAP contributors for Plain and Non-Plain predictions. Only words with at least five positive SHAP occurrences are included. For v5 (XLM-R), class labels were corrected after verifying the reversed raw output mapping.

Model	Class	Rank	Word	Mean SHAP	Occurrences
v5 (XLM-R)	Non-Plain	1	tekintettel [with regard to]	0.59434	11
v5 (XLM-R)	Non-Plain	2	amennyiben [provided that/if]	0.46533	34
v5 (XLM-R)	Non-Plain	3	követően [following/subsequent to]	0.45181	22
v5 (XLM-R)	Non-Plain	4	minősül [qualifies as]	0.43290	24
v5 (XLM-R)	Non-Plain	5	során [during/in the course of]	0.41844	26
v5 (XLM-R)	Non-Plain	6	vonatkozóan [concerning/regarding]	0.40068	10
v5 (XLM-R)	Non-Plain	7	történő [performed/carried out]	0.38234	43
v5 (XLM-R)	Non-Plain	8	esetében [in the case of]	0.37762	28
v5 (XLM-R)	Non-Plain	9	minősülő [qualifying as]	0.35675	7
v5 (XLM-R)	Non-Plain	10	esetén [in case of/upon]	0.33901	51
v5 (XLM-R)	Non-Plain	11	vonatkozásában [in respect of]	0.33376	6
v5 (XLM-R)	Non-Plain	12	részére [to/for]	0.33211	22
v5 (XLM-R)	Non-Plain	13	felé [toward]	0.33117	7
v5 (XLM-R)	Non-Plain	14	értelmében [pursuant to/under]	0.32882	9
v5 (XLM-R)	Non-Plain	15	tekintetében [in terms of/regarding]	0.29395	7
v5 (XLM-R)	Non-Plain	16	lehetősége [possibility of]	0.27940	14
v5 (XLM-R)	Non-Plain	17	került [has been (passive marker)]	0.27697	5
v5 (XLM-R)	Non-Plain	18	mely [which]	0.27559	9
v5 (XLM-R)	Non-Plain	19	foglaltak [provisions/contained in]	0.26431	5
v5 (XLM-R)	Non-Plain	20	való [being]	0.26220	14
v5 (XLM-R)	Plain	1	aeo- [AEO—Authorized Economic Operator]	0.14190	20
v5 (XLM-R)	Plain	2	fémkereskedelmiengedély- [metal trade license]	0.11539	6
v5 (XLM-R)	Plain	3	eu- [EU]	0.11106	7
v5 (XLM-R)	Plain	4	rex- [REX—Registered Exporter]	0.11103	5
v5 (XLM-R)	Plain	5	héa- [VAT—Value Added Tax]	0.11093	5
v5 (XLM-R)	Plain	6	xml- [XML]	0.10877	9
v5 (XLM-R)	Plain	7	szja- [PIT—Personal Income Tax]	0.10396	6
v5 (XLM-R)	Plain	8	ezért [therefore]	0.08832	5
v5 (XLM-R)	Plain	9	nav- [NTCA—Tax Authority]	0.08199	20
v5 (XLM-R)	Plain	10	oss- [OSS—One Stop Shop]	0.08044	8
v5 (XLM-R)	Plain	11	nav [NTCA]	0.07728	91
v5 (XLM-R)	Plain	12	nincs [there is no]	0.07274	9
v5 (XLM-R)	Plain	13	ilyenkor [in this case]	0.07016	8
v5 (XLM-R)	Plain	14	jelszó [password]	0.07011	6
v5 (XLM-R)	Plain	15	vállalkozónak [for the entrepreneur]	0.06777	6
v5 (XLM-R)	Plain	16	több [more/multiple]	0.05458	9
v5 (XLM-R)	Plain	17	például [for example]	0.05246	19
v5 (XLM-R)	Plain	18	is [also/as well]	0.04766	90
v5 (XLM-R)	Plain	19	utáni [after/post-]	0.04662	7
v5 (XLM-R)	Plain	20	még [still/yet]	0.04652	7
v1 (huBERT)	Plain	1	aeo- [AEO]	0.19480	18
v1 (huBERT)	Plain	2	oss- [OSS]	0.13371	7
v1 (huBERT)	Plain	3	eáfa [e-VAT]	0.13210	7
v1 (huBERT)	Plain	4	nav- [NTCA]	0.10798	18
v1 (huBERT)	Plain	5	ilyenkor [in this case]	0.10262	8

Table 9. *Cont.*

Model	Class	Rank	Word	Mean SHAP	Occurrences
v1 (huBERT)	Plain	6	éig [until the year]	0.09403	6
v1 (huBERT)	Plain	7	szja- [PIT]	0.08554	6
v1 (huBERT)	Plain	8	gépjárműadót [motor vehicle tax]	0.08200	8
v1 (huBERT)	Plain	9	bevallásban [in the tax return]	0.08173	5
v1 (huBERT)	Plain	10	áfa [VAT]	0.07633	9
v1 (huBERT)	Plain	11	szja [PIT]	0.07468	18
v1 (huBERT)	Plain	12	fontos [important]	0.07291	5
v1 (huBERT)	Plain	13	kifizető [payer/employer]	0.07231	6
v1 (huBERT)	Plain	14	ányk [electronic tax form wizard]	0.06684	5
v1 (huBERT)	Plain	15	számlán [on the invoice]	0.06645	6
v1 (huBERT)	Plain	16	nav [NTCA]	0.06457	85
v1 (huBERT)	Plain	17	olvashat [can read]	0.06294	6
v1 (huBERT)	Plain	18	című [entitled/titled]	0.05962	7
v1 (huBERT)	Plain	19	kiküldetési [official mission/posting]	0.05730	5
v1 (huBERT)	Plain	20	feltétel [condition/requirement]	0.05582	5
v1 (huBERT)	Non-Plain	1	tekintettel [with regard to]	0.40584	11
v1 (huBERT)	Non-Plain	2	során [during/in the course of]	0.31597	25
v1 (huBERT)	Non-Plain	3	amennyiben [provided that/if]	0.31117	34
v1 (huBERT)	Non-Plain	4	esetében [in the case of]	0.27784	30
v1 (huBERT)	Non-Plain	5	történő [performed/carried out]	0.26837	43
v1 (huBERT)	Non-Plain	6	vonatkozásában [in respect of]	0.25000	6
v1 (huBERT)	Non-Plain	7	minősül [qualifies as]	0.24914	22
v1 (huBERT)	Non-Plain	8	követően [following]	0.24495	24
v1 (huBERT)	Non-Plain	9	figyelemmel [with attention to]	0.24079	5
v1 (huBERT)	Non-Plain	10	részére [to/for]	0.23458	21
v1 (huBERT)	Non-Plain	11	számára [for (someone)]	0.22804	11
v1 (huBERT)	Non-Plain	12	tekintetében [in terms of]	0.21205	7
v1 (huBERT)	Non-Plain	13	vonatkozóan [concerning]	0.19195	10
v1 (huBERT)	Non-Plain	14	esetén [in case of]	0.18471	49
v1 (huBERT)	Non-Plain	15	értelmében [pursuant to/under]	0.18148	9
v1 (huBERT)	Non-Plain	16	stb [etc.]	0.17422	7
v1 (huBERT)	Non-Plain	17	esetben [in case]	0.17068	28
v1 (huBERT)	Non-Plain	18	tekinthető [can be considered]	0.16399	6
v1 (huBERT)	Non-Plain	19	pl [e.g.,]	0.15701	10
v1 (huBERT)	Non-Plain	20	alábbi [following/below]	0.15576	17

At the same time, the Plain-oriented lists contain several domain-specific or institutional terms, such as *NAV*, *SZJA*, *ÁFA*, *AEO*, and *OSS*. This does not mean that these terms are inherently plain. Rather, it reflects the nature of the corpus. Plain Language in this domain does not require the elimination of legal or tax administrative terminology, but the use of such terminology within clearer and more reader-oriented sentence structures. In other words, the models appear to distinguish not only between technical and non-technical vocabulary, but also between different ways in which technical vocabulary is embedded in administrative prose.

These findings refine the interpretation of the local SHAP examples. The models do not rely exclusively on isolated lexical shortcuts, nor do they purely capture abstract syntactic complexity. Instead, they combine register-sensitive linguistic cues with corpus-specific lexical signals. This mixed behavior is especially important for deployment. The models can identify recurring markers of legal–administrative complexity, but their predictions may also remain sensitive to the taxation-specific terminology present in the training data.

7.6. Limitations

While our findings offer strong support for fine-tuned multilingual transformers in Hungarian PL classification, several limitations must be acknowledged to contextualize the results and guide future improvements.

The size and scope of the training corpus present inherent constraints. Although sentence-level alignment and preprocessing ensured a high-quality dataset, the overall volume remains modest by modern deep learning standards. This restricts the capacity of very large models to generalize reliably and increases the risk of overfitting, particularly in high-parameter architectures such as XLM-RoBERTa-large; while machine-translated English augmentations partially mitigated data sparsity, they may introduce noise due to mismatches in syntax, terminology, or stylistic register. This risk is especially relevant in the legal domain, where small linguistic shifts can alter semantic intent.

The binary classification framework adopted in this study may oversimplify the inherently gradient nature of linguistic clarity. Sentences do not always fall neatly into either "Plain" or "Non-Plain" categories. Intermediate cases are common and often context-dependent. Although dichotomous labels facilitate supervised learning and evaluation, they may obscure fine-grained stylistic or structural improvements. Future work could explore regression or ordinal classification schemes to better model the continuum of clarity.

In a real-time writing assistant, the classifier should not be interpreted as providing a definitive legal or linguistic judgment. Instead, its output could be used to rank sentences according to revision priority, with the decision threshold adjusted to favor higher recall for potentially Non-Plain sentences. Cases with scores close to the decision boundary should be referred for human review rather than automatically flagged as requiring revision. The system should also avoid automatic rewriting of legally sensitive content, and domain-specific terminology alone should not be treated as sufficient evidence of Non-Plainness.

Our evaluation framework relies on conventional classification metrics (precision, recall, F1-score), which, although useful for benchmarking, may not fully reflect practical value in legal drafting. A classifier may correctly flag structurally complex sentences while failing to identify opaque but syntactically simple constructions. Furthermore, high aggregate scores may mask the misclassification of particularly salient or high-stakes clauses. Incorporating human-centered metrics, such as revision effort, reading time, or user-rated clarity, could provide richer insight into real-world applicability.

Limitations also arise from the domain-specific nature of the corpus. All texts originate from the Hungarian Tax and Customs Administration (NAV), representing a narrow segment of bureaucratic legal discourse. The degree to which these results generalize to other legal subdomains, e.g., judicial decisions, legislative texts, or contractual language remains uncertain. Domain-adaptive pretraining or few-shot learning may be necessary to accommodate stylistic variation across genres. The aggregated SHAP analysis further supports this domain-related limitation; while Non-Plain predictions were strongly associated with linguistically interpretable administrative formulae, Plain predictions also showed sensitivity to NAV-specific tax and institutional terminology. This indicates that part of the models' decision making is tied to the lexical distribution of the source corpus. Consequently, cross-domain use in statutory legislation, court judgments, or contractual texts would require additional validation and, ideally, domain-adaptive data.

We also acknowledge potential label inconsistencies arising from the corpus construction process. Since the dataset was built from sentence-level edits, each change was assumed to reflect a move toward greater clarity. However, editorial interventions may not always have been driven solely by PL criteria. Some edits might reflect structural corrections, content clarification, or stylistic preferences unrelated to PL. This introduces a degree of annotation noise that could affect model learning and explain certain classification

errors, particularly false positives involving domain-specific vocabulary. Future corpus design should therefore include annotation protocols that distinguish PL-driven edits from broader revisions.

Finally, practical deployment raises concerns about reproducibility, transparency, and infrastructure compatibility. Proprietary models like GPT-4o-mini and Gemini 1.0 Pro, while achieving respectable results, are accessible only via third-party APIs, limiting long-term usability, auditability, and control. Even open-weight models become dependent on specific hardware and software stacks once fine-tuned. Ensuring model versioning, containerized environments, and accessible documentation is essential for reproducibility and institutional trust.

Recognizing these limitations provides an important context for the study's findings and outlines actionable directions to build more robust, generalizable, and ethically deployable PL classification systems in legal and administrative domains.

8. Conclusions

This study explored the feasibility of applying transformer-based models for sentence-level classification of legal and administrative texts in Hungarian according to PL standards. Addressing a critical resource gap in Hungarian NLP, we benchmarked zero-shot and fine-tuned variants across monolingual and multilingual architectures, leveraging translation-based data augmentation to enhance performance in a low-resource legal setting.

Our results demonstrate that fine-tuning multilingual models (particularly XLM-RoBERTa-large trained with Hungarian–English parallel data) yields substantial improvements in classification accuracy. The best-performing configuration (v5a) achieved a macro-average F1-score of 0.79, significantly outperforming both monolingual baselines and advanced zero-shot prompting with GPT-4o. This improvement is also meaningful in relation to earlier lightweight supervised baselines on the same corpus family, since the best fine-tuned transformer configuration significantly outperformed both zero-shot GPT-4o and previously published TF-IDF + SVM and fastText approaches. This supports prior evidence that task-specific adaptation and cross-lingual exposure can greatly benefit classification tasks in under-resourced languages.

We also evaluated deployment-relevant dimensions such as model size, inference time, and accessibility. Although the largest models generally achieved high performance, our results show that mid-sized open-weight architectures (particularly XLM-RoBERTa-base with augmented data) can match or even surpass larger models when evaluated on the original Hungarian texts. Combined with their lower latency and infrastructure requirements, these models offer a compelling trade-off between effectiveness and practical constraints such as reproducibility, regulatory compliance, and deployment feasibility in institutional settings.

Importantly, our work illustrates that translation-based data augmentation can meaningfully expand training coverage and improve generalization, especially when applied during training rather than evaluation. This strategy is particularly useful for domains like legal drafting, where annotated corpora are costly and stylistic variation is high.

Beyond aggregate metrics, we applied SHAP-based local interpretability techniques to examine the behavior of the model on both true and false classifications. The analysis revealed that the best-performing model reliably identified linguistic features associated with syntactic complexity and nominalization. At the same time, SHAP helped uncover limitations in model generalization—such as false positives triggered by frequent domain-specific expressions such as institutional acronyms or formal terminology. These insights not only confirmed the linguistic validity of the model's decision boundaries but also

revealed sources of bias and annotation noise in the corpus, underscoring the value of interpretability tools in model auditing and refinement.

Ultimately, the findings lay the foundation for the building of automated Plain Language assistance tools that help legal professionals and public institutions identify inaccessible language in real time. In practical deployment, such a system should function as a human-in-the-loop prioritization tool rather than as an automated authority on legal clarity. By integrating such models into document drafting workflows, we can support more transparent and citizen-friendly communication, thereby advancing the broader goal of Access to Justice.

Future directions include extending the framework to graded classification (e.g., clarity scoring), adapting the models to other public-facing domains such as healthcare or education, and integrating human-in-the-loop systems for collaborative editing. Additionally, refining interpretability and bias-detection techniques will be essential for ensuring that Plain Language tools are not only accurate but also trustworthy in sensitive legal contexts.

Funding: This research received no external funding. The APC was funded by the author.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study are not publicly available due to copyright restrictions. They may be made available by the corresponding author upon reasonable request, strictly for research purposes and subject to applicable copyright restrictions and permission from the original data provider. The fine-tuned open-weight model checkpoints developed and evaluated in this study are publicly available on Hugging Face: https://huggingface.co/uvegesistvan/huBERTPlain (accessed on 30 May 2026), https://huggingface.co/uvegesistvan/Hun_RoBERTa_base_Plain (accessed on 30 May 2026), https://huggingface.co/uvegesistvan/Hun_RoBERTa_large_Plain (accessed on 30 May 2026), https://huggingface.co/uvegesistvan/Hun_Eng_RoBERTa_base_Plain (accessed on 30 May 2026), and https://huggingface.co/uvegesistvan/Hun_Eng_RoBERTa_large_Plain (accessed on 30 May 2026).

Acknowledgments: The author would like to thank the Plain Language Group of the National Tax and Customs Administration of Hungary (NAV) for providing the original data used in this study. During the preparation of this manuscript, the author used ChatGPT with the GPT-4o and GPT-5.5 models by OpenAI for spell checking and grammar checking. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: Author István Üveges was employed by Griffsoft Zrt. The author declares that the research was conducted in the absence of any other commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Hyland-Wood, B.; Gardner, J.; Leask, J.; Ecker, U.K.H. Toward effective government communication strategies in the era of COVID-19. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 30. [CrossRef]
2. Cappelletti, M.; Garth, B. *Access to Justice*; Sijthoff and Noordhoff: Alphen aan den Rijn, The Netherlands, 1978; Volumes 1–4.
3. Tiersma, P.M. *Legal Language*; University of Chicago Press: Chicago, IL, USA, 1999.
4. Kimble, J. *Writing for Dollars, Writing to Please: The Case for Plain Language in Business, Government, and Law*; Carolina Academic Press: Durham, NC, USA, 2012.
5. Stephens, C.M. Access to Justice Requires Plain Language. *Mich. Bar J.* **2020**, *99*, 46–47.
6. United Nations. Access to Justice Critical in Ensuring Rule of Law, Speakers Stress as Sixth Committee Continues Deliberations on Principle. 2014. Available online: https://press.un.org/en/2014/gal3478.doc.htm (accessed on 30 May 2026).
7. Vinnai, E. Az első "jog és nyelv" kutatás hazánkban. *Alkalm. Nyelvészeti Közlemények* **2014**, *9*, 60–67.
8. Mazur, B. Revisiting Plain Language. *Tech. Commun.* **2000**, *47*, 205–215.
9. *ISO 24495-1:2023*; Plain Language—Part 1: Governing Principles and Guidelines. International Organization for Standardization: Geneva, Switzerland, 2023.

10. Orosz, G.; Szabó, G.; Berkecz, P.; Szántó, Z.; Farkas, R. Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines. In *Text, Speech, and Dialogue*; Ekštein, K., Pártl, F., Konopík, M., Eds.; Springer Nature: Cham, Switzerland, 2023; pp. 58–69.
11. Prótár, N.; Nemeskey, D.M. huPWKP: A Hungarian Text Simplification Corpus. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, Varna, Bulgaria, 4–6 September 2023* ; Mitkov, R., Angelova, G., Eds.; INCOMA Ltd.: Shoumen, Bulgaria, 2023; pp. 898–907. [CrossRef] [PubMed]
12. Zhu, Z.; Bernhard, D.; Gurevych, I. A Monolingual Tree-based Translation Approach to Sentence Simplification. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, 23–27 August 2010.
13. Üveges, I. Comprehensibility and Automation: Plain Language in the Era of Digitalization. *TalTech J. Eur. Stud.* **2022**, *12*, 64–86. [CrossRef]
14. Üveges, I. A jogi szövegek közérthetösége: Modern nyelvfeldolgozás és szoftveres támogatás. *Magy. Jogi Nyelv* **2024**, *8*, 33–40.
15. Ravinetto, R.; Singh, J.A. Responsible dissemination of health and medical research: Some guidance points. *BMJ Evid. Based Med.* **2023**, *28*, 144–147. [CrossRef] [PubMed]
16. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
17. Schriver, K.A. Plain Language in the US Gains Momentum: 1940–2015. *IEEE Trans. Prof. Commun.* **2017**, *60*, 343–383. [CrossRef]
18. Dale, E.; Chall, J.S. A Formula for Predicting Readability: Instructions. *Educ. Res. Bull.* **1948**, *27*, 37–54.
19. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233. [CrossRef] [PubMed]
20. DuBay, W.H. The Principles of Readability. 2004. Available online: https://files.eric.ed.gov/fulltext/ED490073.pdf (accessed on 30 May 2026).
21. Farrall, M.L. *Reading Assessment: Linking Language, Literacy, and Cognition*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
22. Crossley, S.A.; Allen, D.B.; McNamara, D.S. Text readability and intuitive simplification: A comparison of readability formulas. *Read. Foreign Lang.* **2011**, *23*, 84–101. [CrossRef]
23. Redish, J.C. Readability formulas have even more limitations than Klare discusses. *ACM J. Comput. Doc.* **2000**, *24*, 132–137. [CrossRef]
24. Tekfi, C. Readability Formulas: An Overview. *J. Doc.* **1987**, *43*, 261–273. [CrossRef]
25. Byrd, D.; Mintz, T.H. *Discovering Speech, Words, and Mind*; Wiley-Blackwell: Chichester, UK, 2010.
26. Friederici, A.D. Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* **2002**, *6*, 78–84. [CrossRef] [PubMed]
27. Gibson, E.; Pearlmutter, N.J. Constraints on sentence comprehension. *Trends Cogn. Sci.* **1998**, *2*, 262–268. [CrossRef] [PubMed]
28. Lukács, A.; Pléh, C. (Eds.) *Pszicholingvisztika: Magyar Pszicholingvisztikai Kézikönyv*; Akadémiai Kiadó: Budapest, Hungary, 2014.
29. Charrow, R.P.; Charrow, V.R. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Rev.* **1979**, *79*, 1306–1374. [CrossRef]
30. Kas, B.; Lukács, A. Processing relative clauses by Hungarian typically developing children. *Lang. Cogn. Process.* **2012**, *27*, 500–538. [CrossRef] [PubMed]
31. Pléh, C. *A Lélek és a Nyelv*; Akadémiai Kiadó: Budapest, Hungary, 2015. [CrossRef]
32. Asprey, M.M. *Plain Language for Lawyers*; Federation Press: Sydney, Australia, 2003.
33. Garner, B.A. *Legal Writing in Plain English: A Text with Exercises*; University of Chicago Press: Chicago, IL, USA, 2001.
34. Mellinkoff, D. *The Language of the Law*; Resource Publications: Eugene, OR, USA, 2004.
35. Cutts, M. *The Plain English Guide*; Oxford University Press: Oxford, UK, 1996.
36. Cutts, M. *Oxford Guide to Plain English*, 5th ed.; Oxford University Press: Oxford, UK, 2020.
37. Willerton, R. *Plain Language and Ethical Action: A Dialogic Approach to Technical Content in the 21st Century*; Routledge: New York, NY, USA, 2015. [CrossRef]
38. Wydick, R.C. *Plain English for Lawyers*, 5th ed.; Carolina Academic Press: Durham, NC, USA, 2005.
39. Szabó, M.; Vinnai, E. *A Törvény Szavai*; Bíbor Kiadó: Miskolc, Hungary, 2018.
40. Balogh, D. Alanyi szerkezetek a magyar jogi nyelvben. In *A Törvény Szavai: Az OTKA-112172 Kutatási Zárókonferencia Anyaga Miskolc, ME—MAB, 2018 Május 25*; Szabó, M., Vinnai, E., Eds.; Bíbor Kiadó: Miskolc, Hungary, 2018; pp. 71–94.
41. Dobos, C. Állítmányi szerkezetek a magyar jogi nyelvben. In *A Törvény Szavai: Az OTKA-112172 Kutatási Zárókonferencia Anyaga Miskolc, ME—MAB, 2018 Május 25*; Szabó, M., Vinnai, E., Eds.; Bíbor Kiadó: Miskolc, Hungary, 2018; pp. 37–69.
42. Goodrich, P. *Legal Discourse*; Palgrave Macmillan: London, UK, 1987. [CrossRef]
43. Levi, J.N.; Walker, A.G. (Eds.) *Language in the Judicial Process*; Springer: Boston, MA, USA, 1990. [CrossRef]
44. Loftus, E.F.; Palmer, J.C. Reconstruction of automobile destruction: An example of the interaction between language and memory. *J. Verbal Learn. Verbal Behav.* **1974**, *13*, 585–589. [CrossRef]
45. Vincze, V. A Miskolci Jogi Korpusz nyelvi jellemzöi. In *A Törvény Szavai: Az OTKA-112172 Kutatási Zárókonferencia Anyaga Miskolc, ME—MAB, 2018 Május 25*; Szabó, M., Vinnai, E., Eds.; Bíbor Kiadó: Miskolc, Hungary, 2018; pp. 9–36.
46. Vinnai, E. *Jog és Nyelv Határán: A Jogi Nyelvhasználat Nemzetközi és Hazai Kutatása*; Gondolat: Budapest, Hungary, 2017.

47. Kálmán, R.; Márki, D.; Németh, K.; Szabó, E.; Szecskó, E.; Tóth, J.; Tribl, N. *Közjogi Fogalmak Közérthetően*; Iurisperitus Kiadó: Szeged, Hungary, 2020.
48. Kim, J.; Leijnen, S.; Beinborn, L. Considering Human Interaction and Variability in Automatic Text Simplification. In Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Miami, FL, USA, 15 November 2024. [CrossRef]
49. Stodden, R. EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages. In Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Miami, FL, USA, 15 November 2024. [CrossRef]
50. Yasseri, T.; Kornai, A.; Kertész, J. A practical approach to language complexity: A Wikipedia case study. *PLoS ONE* **2012**, *7*, e33379. [CrossRef] [PubMed]
51. Pottmann, D.M. Leichte Sprache and Einfache Sprache–German Plain Language and teaching DaF German as a foreign language. *Stud. Linguist.* **2020**, *38*, 81–94. [CrossRef]
52. Radünzel, C. Leichte Sprache: Eine linguistische Betrachtung eines neuen sprachlichen Phänomens auf der Grundlage polnischer und deutscher Beispieltexte. *Z. Slaw.* **2017**, *62*, 48–95. [CrossRef]
53. Xu, W.; Callison-Burch, C.; Napoles, C. Problems in Current Text Simplification Research: New Data Can Help. *Trans. Assoc. Comput. Linguist.* **2015**, *3*, 283–297. [CrossRef]
54. Alva-Manchego, F.; Martin, L.; Bordes, A.; Scarton, C.; Sagot, B.; Specia, L. ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020. [CrossRef]
55. Üveges, I. Közérthetőség és Automatizáció-Kísérletek a Jog, Természetesnyelv-Feldolgozás és Informatika Határán. Ph.D. Thesis, Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola, Szeged, Hungary, 2024. [CrossRef]
56. Chandrasekar, R.; Doran, C.; Srinivas, B. Motivations and Methods for Text Simplification. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 5–9 August 1996.
57. Siddharthan, A. A survey of research on text simplification. *Int. J. Appl. Linguist.* **2014**, *165*, 259–298. [CrossRef]
58. Dong, Y.; Li, Z.; Rezagholizadeh, M.; Cheung, J.C.K. EditNTS: A Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing. *arXiv* **2019**, arXiv:1906.08104.
59. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*; Curran Associates, Inc.: Red Hook, NY, USA, 2014.
60. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
61. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019.
62. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
63. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 845–860. [CrossRef]
64. Li, Z.; Belkadi, S.; Micheletti, N.; Han, L.; Shardlow, M.; Nenadic, G. Investigating Large Language Models and Control Mechanisms to Improve Text Readability of Biomedical Abstracts. *arXiv* **2023**, arXiv:2309.13202. [CrossRef]
65. Basu, C.; Vasu, R.; Yasunaga, M.; Yang, Q. Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. *arXiv* **2023**, arXiv:2302.09155. [CrossRef]
66. Zhang, X.; Lapata, M. Sentence Simplification with Deep Reinforcement Learning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017. [CrossRef]
67. National Institutes of Health. Clearly Communicating Research Results Across the Clinical Trials Continuum. 2025. Available online: https://stagetestdomain3.nih.gov/health-information/nih-clinical-research-trials-you/clearly-communicating-research-results-across-clinical-trials-continuum (accessed on 30 May 2026).
68. Guo, Y.; Qiu, W.; Wang, Y.; Cohen, T. Automated Lay Language Summarization of Biomedical Scientific Reviews. *arXiv* **2022**, arXiv:2012.12573.
69. Tribl, N. Az alkotmányos identitás fogalomrendszere jogelméleti megközelítésben. *Jogelméleti Szle.* **2018**, *19*, 151–164. [CrossRef]
70. Vági, R.; Üveges, I. Laws Clearly: Large language models and Plain Language transformation. *Magy. Nyelvőr* **2024**, *148*, 696–716. [CrossRef]
71. OpenAI. GPT-4o System Card. *arXiv* **2024**, arXiv:2410.21276. [CrossRef]

72. Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Fagbohungbe, T.; Orife, I.; Emezue, C.; Dossou, S.; Osei, S.; Adewumi, T.; et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2144–2160.
73. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.
74. Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 6282–6293.
75. Ruder, S. Transfer Learning—Machine Learning's Next Frontier. 21 March 2017. Available online: https://www.ruder.io/transfer-learning/ (accessed on 30 May 2026).
76. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Do not Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
77. Nemeskey, D.M. Introducing huBERT. In *Proceedings of the XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021), Szeged, Hungary, 28–29 January 2021*; Berend, G., Gosztolya, G., Vincze, V., Eds.; University of Szeged, Faculty of Science and Informatics, Institute of Informatics: Szeged, Hungary, 2021; pp. 3–14. Available online: https://acta.bibl.u-szeged.hu/73353/1/msznykonf_017_003-014.pdf (accessed on 30 May 2026).
78. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
79. OpenAI. GPT-4o mini. 2024. Available online: https://platform.openai.com/docs/models/gpt-4o-mini (accessed on 20 May 2025).
80. Google Cloud. Gemini on Vertex AI Expands. 15 February 2024. Available online: https://cloud.google.com/blog/products/ai-machine-learning/gemini-on-vertex-ai-expands (accessed on 30 May 2026). https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro (accessed on 20 May 2025).
81. Fadaee, M.; Bisazza, A.; Monz, C. Data Augmentation for Low-Resource Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017.
82. Artetxe, M.; Schwenk, H. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 597–610. [CrossRef]
83. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020.
84. Lakew, S.M.; Negri, M.; Turchi, M. Low Resource Neural Machine Translation: A Benchmark for Five African Languages. *arXiv* **2020**, arXiv:2003.14402.
85. Radford, A.; Wu, J.; Amodei, D.; Clark, J.; Brundage, M.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *Openai Blog* **2019**, *1*, 9.
86. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
87. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training compute-optimal large language models. *arXiv* **2022**, arXiv:2203.15556.
88. Butt, M. The Light Verb Jungle: Still Hacking Away. In *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*; Amberber, M., Harvey, M., Baker, B., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 48–78.