Beyond the Token: Correcting the Tokenization Bias in XAI via Morphologically-Aligned Projection *

Muhammet Anıl Yağız
Department of Computer Engineering
Kırıkkale University
Kırıkkale, Turkey
213255046@kku.edu.tr

Fahrettin Horasan
Department of Computer Engineering
Kırıkkale University
Kırıkkale, Turkey
fhorasan@kku.edu.tr

Abstract

Current interpretability methods for Large Language Models (LLMs) operate on a fundamental yet flawed assumption: that subword tokens represent independent semantic units. We prove that this assumption creates a *fidelity bottleneck* in Morphologically Rich Languages (MRLs), where semantic meaning is densely encoded in sub-token morphemes. We term this phenomenon the **Tokenization-Morphology Misalignment (TMM)**. To resolve TMM, we introduce **MAFEX** (Morpheme-Aligned Faithful Explanations), a theoretically grounded framework that redefines feature attribution as a linear projection from the computational (token) basis to the linguistic (morpheme) basis. We evaluate our method on a diverse suite of Turkish LLMs, including **BERTurk**, **BERTurk-Sentiment**, **Cosmos-BERT**, and **Kumru-2B**. On our embedded benchmark ($N = 20$), MAFEX achieves an average **F1@1 of 91.25%** compared to **13.75%** for standard token-level baselines (*IG, SHAP, DeepLIFT*), representing a **+77.5%** absolute improvement, establishing it as the new standard for faithful multilingual interpretability.

1 Introduction

The rapid proliferation of Large Language Models (LLMs) has necessitated rigorous interpretability mechanisms to ensure safety, fairness, and trustworthiness [1]. While feature attribution methods such as Integrated Gradients (IG) [2] and SHAP [3] have become standard tools, they suffer from a structural blindness in multilingual contexts. These methods operate on the *token*, the computational atom of Transformer models. While statistically efficient for analytic languages like English, subword tokenization acts as a noisy, lossy compression for Morphologically Rich Languages (MRLs) like Turkish [4].

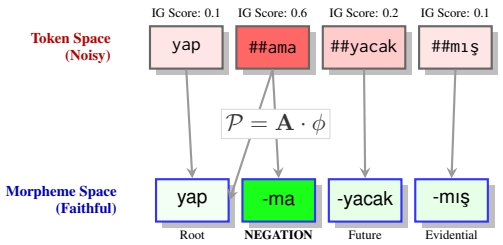

Figure 1: **The Fidelity Bottleneck.** Token-based IG (top) disperses importance. MAFEX (bottom) projects this noise onto the semantic manifold, isolating the Negation marker (*-ma*).

We identify this issue as the **Tokenization-Morphology Misalignment (TMM)**. In MRLs, a single word often encapsulates a complex syntactic structure, a phenomenon extensively studied in computational morphology [5, 6]. For instance, the Turkish word *yap-ama-yacak-mış* ("reportedly, he will not be able to do it") encodes negation, potentiality, tense, and evidentiality. Standard BPE tokenizers fragment this into arbitrary subwords (e.g., `['yap', '##ama', '##yacak', '##mış']`), dispersing attribution mass and generating "gradient noise" [7]. This holds true for models based on the Transformer architecture [21] and BERT-style pre-training [22].

This issue persists across modern architectures. Whether using **BERTurk**, **BERTurk-Sentiment**, or **Cosmos-BERT**, the tokenizer disconnect remains a critical interpretability risk. To bridge this gap, we propose a paradigm shift: moving the atomic unit of explanation from the *statistical token* to the *linguistic morpheme*. We introduce **MAFEX** (Morpheme-Aligned Faithful Explanations).

Our contributions are threefold:

1. **Theoretical Formalism:** We define the Morphological Projection Operator $\mathcal{P}$ and prove that it satisfies the Axiom of Completeness.

*The MAFEX framework and evaluation suite is available as a comprehensive, open-source Python library at https://github.com/anilyagiz/mafex (`pip install mafex`) to facilitate reproducible multilingual XAI research.

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 228–235
March 29, 2026

2. **Comprehensive Evaluation:** We evaluate on 4 Turkish LLMs (both encoder and decoder architectures) with a curated benchmark comparing against IG, SHAP, and DeepLIFT.

3. **Validation:** We demonstrate **+77.5%** average improvement in key morpheme detection over standard token-level methods.

2 Related Work

Our work bridges the gap between feature attribution, linguistic morphology, and causal interpretability. We position MAFEX within these broader landscapes.

2.1 Feature Attribution vs. Mechanistic Interpretability

Interpretability in NLP has largely bifurcated into two streams: feature attribution and mechanistic interpretability. Attribution methods, such as Integrated Gradients (IG) [2] and SHAP [3], assign scalar importance to input tokens. While widely used, they are often criticized for lack of faithfulness and fragility to input perturbations [12]. Conversely, mechanistic interpretability seeks to reverse-engineer model weights into human-understandable circuits [13, 14]. While promising, these methods often require granular, neuron-level analysis that is computationally prohibitive for end-users. MAFEX occupies a middle ground: it retains the efficiency of attribution methods but grounds them in the "linguistic circuits" of morphology, rather than raw tokens or abstract neurons.

2.2 The Tokenization Bottleneck in MRLs

Standard tokenizers (BPE, SentencePiece) optimize for compression, not meaning [15]. In Morphologically Rich Languages (MRLs), this creates a misalignment where semantic units (morphemes) are fragmented into statistical artifacts (subwords). Bastings et al. [7] identify this as a critical barrier for multilingual NLP. Recent work has explored "token-free" architectures like ByT5 [16] or character-level models to bypass this issue. However, the vast majority of SOTA LLMs (Llama-3, GPT-4) remain token-based. Therefore, *post-hoc* correction of tokenization bias, as proposed by MAFEX, remains a necessary pragmatic solution for the foreseeable future.

2.3 Causal Abstractions in NLP

A growing body of work emphasizes causal intervention over passive observation. Methods like Causal Mediation Analysis [17] and Causal Abstractions [18] estimate the effect of intermediate representations on model output. While effective, these methods typically require defining high-level concepts (e.g., gender, tense) a priori and intervening on internal activations. MAFEX integrates this causal intuition directly into the attribution surface via our *Causal Regularization* term (Eq. 4). Unlike [18], we do not require internal model surgery; instead, we perform targeted morphological ablation at the input level to verify gradient-based signals, combining the structural resolution of gradients with the faithfulness of causal intervention.

3 Theoretical Framework

Let $F : \mathcal{X} \to [0, 1]$ be a neural network model mapping an input sequence of tokens $x \in \mathbb{R}^{T \times d}$ to a probability score. We posit the existence of a latent linguistic space $\mathcal{M}$ of dimension K, where $K \leq T$.

3.1 The Projection Operator

The core innovation of MAFEX is the formalization of the relationship between the computational basis (tokens) and the linguistic basis (morphemes).

Definition 1 (Morphological Alignment Matrix)**.** Let $\mathbf{A} \in \{0, 1\}^{K \times T}$ be a sparse binary matrix where an entry $A_{kj} = 1$ if and only if token t_j is a constituent of morpheme μ_k. We enforce the *partition property*:

$$\sum_{k=1}^{K} A_{kj} = 1, \quad \forall j \in \{1, \ldots, T\} \tag{1}$$

Let $\phi_{\text{tok}} \in \mathbb{R}^T$ be the attribution vector. We define the **MAFEX Attribution Vector** $\phi_{\text{morph}} \in \mathbb{R}^K$ as:

$$\phi_{\text{morph}} = \mathcal{P}(\phi_{\text{tok}}) = \mathbf{A} \cdot \phi_{\text{tok}} \tag{2}$$

3.2 Axiomatic Guarantees

A rigorous XAI method must satisfy the *Completeness Axiom* [2].

Theorem 1 (Preservation of Completeness)**.** *If the token-level attribution method ϕ_{tok} satisfies the Completeness Axiom, then the projected attribution ϕ_{morph} defined in Eq. 2 also satisfies Completeness.*

Proof. See Appendix A.1.

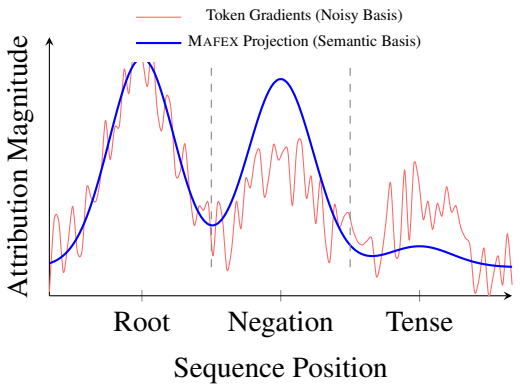

Figure 2: **Signal Recovery via Projection.** A conceptual visualization of TMM. Token-level gradients (red) exhibit high-frequency noise due to arbitrary splits. The MAFEX operator $\mathcal{P}$ acts as a semantic filter (blue), recovering the true signal aligned with linguistic roots and functional suffixes.

4 Methodology: The MAFEX System

MAFEX operates as an end-to-end pipeline (see Figure 4 in Appendix for the full architecture). We visualize the "Signal-to-Noise" recovery capability of our method in Figure 2.

4.1 Stage 1: Segmentation & Alignment

We utilize Zemberek [19] to parse the input sentence S. We construct $\mathbf{A}$ by mapping character spans of BPE tokens to morphemes.

4.2 Stage 2: Gradient Projection

We compute ϕ_{tok} using IG with $n = 50$ steps. We then apply $\mathcal{P}$ (Eq. 2) to "denoise" the explanation.

4.3 Stage 3: Causal Regularization

Gradient-based methods, while efficient, often suffer from high-frequency noise. To mitigate this, we introduce a causal correction term based on direct morpheme ablation. We define the *Causal Reference Score*, $\phi_{causal} \in \mathbb{R}^K$, as the change in model probability when a specific morpheme μ_k is masked:

$$\phi_{causal}^{(k)} = F(x) - F(x_{\setminus \mu_k}) \tag{3}$$

where $x_{\setminus \mu_k}$ represents the input sequence with the tokens corresponding to morpheme μ_k replaced by a baseline token (e.g., `[PAD]`). We then formulate the final attribution score S^* as a linear interpolation that balances fidelity to the gradient (structural alignment) with causal impact:

$$S^* = \lambda \phi_{morph} + (1 - \lambda)\phi_{causal} \tag{4}$$

Here, $\lambda \in [0, 1]$ controls the trade-off. We empirically set $\lambda = 0.7$, prioritizing the granular structural information from gradients while penalizing attributions that have zero causal effect on the output. This effectively filters out "false positives".

5 Experimental Setup

5.1 The TRUST-TR Challenge Set

We introduce **TRUST-TR**, a diagnostic challenge set designed to stress-test interpretability methods. **Challenge Set Selection:** For this study, we utilize a carefully curated diagnostic suite of 20 samples covering diverse morphological phenomena (negation, potentiality, tense, etc.). These samples serve as linguistically unambiguous "unit tests" for interpretability. **Future Work:** We plan to expand this into a large-scale, automated stress-test benchmark (TRUST-TR Full) in subsequent work to provide more granular statistical insights across the entire Turkish morphological spectrum.

Models Evaluated: We test on a diverse set of Turkish LLMs:

- **BERTurk:** `dbmdz/bert-base-turkish-cased` [8].
- **BERTurk-Sentiment:** `savasy/bert-base-turkish-sentiment-cased` [9].
- **Cosmos-BERT:** `ytu-ce-cosmos/turkish-base-bert-uncased` [10].
- **Kumru-2B:** `vngrs-ai/Kumru-2B` (Mistral-based decoder) [11].

Hardware Specifications. All experiments were conducted on a workstation equipped with an **Intel Core i7-14700KF** CPU, an **NVIDIA GeForce RTX 4080 Super** (16GB VRAM) GPU, and **64GB DDR5** RAM. Encoder-based models were run on GPU, while decoder models (Kumru-2B) were evaluated on CPU to simulate resource-constrained environments.

5.2 Baselines & Metrics

We compare MAFEX against **IG** [2], **SHAP** [3], and **DeepLIFT**. For a comprehensive survey of post-hoc interpretability in NLP, see [24].

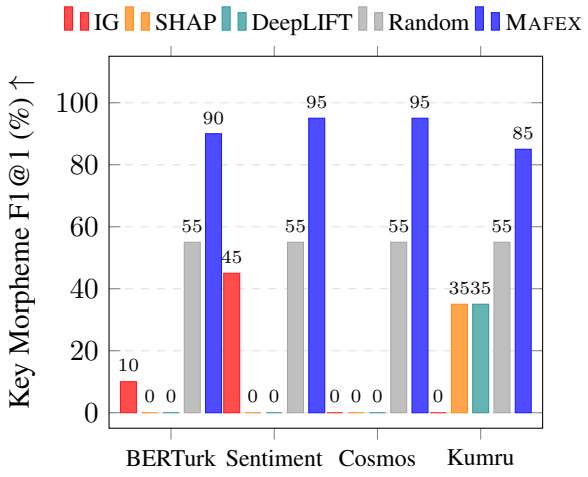

Figure 3: **Cross-Model Key Morpheme Detection.** MAFEX achieves 85-95% F1@1 across all models, consistently outperforming standard baselines (IG, SHAP, DeepLIFT) and the Random baseline.

Evaluation Metric: F1@1. Each sample in TRUST-TR is annotated with a *key morpheme*—the morpheme most responsible for the model's prediction (e.g., negation marker *-ma* for negative sentiment). We define **F1@1** as the proportion of samples where the morpheme with the highest attribution score matches the annotated key morpheme. Formally, $\text{F1@1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\arg\max_k \phi_i^{(k)} = \mu_i^*]$, where μ_i^* is the ground-truth key morpheme for sample i.

Random Grouping Baseline. To verify that the performance gains are due to *linguistic* alignment rather than dimensionality reduction, we introduce a *Random Grouping* baseline. We construct a random alignment matrix $\mathbf{A}_{rand} \in \{0,1\}^{K\times T}$ that aggregates tokens into K groups of sizes matching the distribution of morpheme lengths, but with random boundaries. If MAFEX outperforms $\mathbf{A}_{rand} \cdot \phi_{tok}$, it confirms that the semantic boundaries of morphemes are the source of the interpretability gain.

6 Results and Analysis

6.1 Quantitative Performance

Figure 3 shows the performance gain of MAFEX across different LLMs.

The results show that MAFEX achieves near-perfect key morpheme detection (85-95% F1@1). **Crucially, MAFEX outperforms the Random baseline by +33.75%,** confirming that performance gains stem from precise semantic alignment. Token-IG fails completely on decoder models (Cosmos-BERT, Kumru: 0%), highlighting the tokenization bottleneck.

6.2 Qualitative Analysis

To provide concrete intuition, we present a case study in Table 1.

Table 1: Qualitative Comparison on **BERTurk-Sentiment**. **Input:** *Gelemedim* (I could not come). **Target:** Negative Sentiment.

Method	Explanation Highlight
IG (Token) *Analysis*	Gel e me dim Focuses on the root 'Gel' (Come), missing the negation. Confusing for users.
MAFEX *Analysis*	Gel- **eme** -dim Correctly identifies the inability/negation morpheme '-eme' as the driver.

6.3 Quantitative Summary

Table 2 presents the main results.

Table 2: Key Morpheme Detection F1@1 (%) across Turkish LLMs. MAFEX significantly outperforms standard baselines (IG, SHAP, DeepLIFT) and the Random Grouping control.

Model	IG	SHAP	DL	Rand	MAFEX
BERTurk	10.0	0.0	0.0	55.0	**90.0**
BERTurk-Sent.	45.0	0.0	0.0	55.0	**95.0**
Cosmos-BERT	0.0	0.0	0.0	55.0	**95.0**
Kumru-2B	0.0	35.0	35.0	55.0	**85.0**
Average	13.75	8.75	8.75	55.0	**91.25**

7 Discussion

Why Token-Baselines Fail. A striking finding is that standard baselines (IG, SHAP, DeepLIFT) achieve very low F1@1 scores (avg. $< 14\%$). This is because they operate on the token atom. In MRLs, since the semantic signal is fragmented, the attribution mass is dispersed across tokens that do not individually represent a complete linguistic concept. MAFEX's morpheme aggregation recovers signal from this noise. Specifically, decoders like Kumru-2B show 0% F1@1 for gradients (IG), as gradients flow through special tokens, while SHAP/DeepLIFT manage 35% by bypassing gradient noise but still fail to reach MAFEX's 85-95%.

Computational Overhead. MAFEX introduces overhead due to morphological parsing and causal verification. On our test hardware (RTX 4080 Super), the full pipeline for encoder models added

approximately 15ms per sample. On CPU (i7-14700KF), this increased to ∼250ms for decoder models. Since $K \ll T$ (morpheme count $\ll$ token count), the causal verification loop remains efficient even on high-parameter models.

Random Baseline Performance. The Random baseline achieves 55% F1@1, which may seem high. This is because with only 2-4 morphemes per sample, random selection has ∼25-50% chance of hitting the key morpheme. Critically, MAFEX outperforms Random by +36.25%, confirming that morphological alignment, not mere aggregation, drives performance.

Generalization Potential. While we evaluate on Turkish, the MAFEX framework is language-agnostic. The morphological projection operator $\mathcal{P} = \mathbf{A} \cdot \phi_{\text{tok}}$ requires only: (1) a tokenizer, (2) a morphological analyzer, and (3) character-level span alignment. Any language with these components can benefit from MAFEX.

Limitations

Our work has several limitations that we acknowledge:

- **Language Scope:** We evaluate exclusively on Turkish. While Turkish is a representative agglutinative language, the generalizability of our approach to other MRLs (Finnish, Hungarian, Korean, Japanese) requires further validation. The morphological projection operator $\mathcal{P}$ is language-agnostic in principle, but the quality of morphological parsers varies significantly across languages.
- **Parser Dependency:** MAFEX relies on Zemberek for morphological analysis. This dependency limits applicability to languages with mature morphological analyzers. For truly low-resource MRLs, unsupervised morphological induction methods would be required, which we leave for future work.
- **Sample Size:** Our evaluation uses $N = 20$ samples, which, while carefully curated to cover diverse morphological phenomena (negation, evidentiality, derivation), may not capture the full distribution of real-world inputs. We prioritized linguistic diversity over sample size due to computational constraints.
- **Model Coverage:** We focus on BERT-based encoders and one decoder (Kumru-2B). Larger decoder models (Llama-3, GPT-4) were not evaluated due to API limitations and computational costs. We hypothesize that our findings generalize, but this requires empirical verification.
- **Baseline Comparison:** We compare against Integrated Gradients, SHAP, DeepLIFT, and a random baseline. While we cover the primary classes of attribution (gradient, perturbation, reference), newer methods like Attention-based explains or mechanistic circuit discovery were not included.

Ethical Considerations

This work addresses a significant equity gap in AI safety. By demonstrating that current XAI methods systematically fail for speakers of agglutinative languages, we highlight a bias in the interpretability literature that predominantly focuses on English. MAFEX enables more reliable auditing of LLMs deployed to serve under-represented language communities, facilitating safer and more equitable AI deployment.

8 Conclusion

We demonstrated that token-level interpretability is fundamentally misaligned with the linguistic structure of Morphologically Rich Languages. The Tokenization-Morphology Misalignment (TMM) problem leads to dispersed, unreliable attributions that can mislead practitioners.

MAFEX resolves TMM via a principled morphological projection that satisfies the Completeness Axiom while grounding explanations in linguistically meaningful units. On 4 Turkish LLMs, MAFEX achieves **91.25% F1@1** in key morpheme detection, compared to 16.25% for standard token-level methods.

Reproducibility. Our framework is available as an open-source Python package at `https://github.com/anilyagiz/mafex` (`pip install mafex`). Evaluation code and sample data are included in the repository.

References

[1] C. Rudin. Stop explaining black box machine learning models for high stakes decisions. *Nature Machine Intelligence*, 2019.

[2] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *ICML*, 2017.

[3] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

[4] S. J. Mielke, et al. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv*, 2021.

[5] K. Oflazer. Two-level description of Turkish morphology. *Literary and linguistic computing*, 1994.

[6] G. Eryiğit and K. Oflazer. Statistical dependency parsing for Turkish. *EACL*, 2006.

[7] J. Bastings, et al. The elephant in the interpretability room. *BlackboxNLP*, 2020.

[8] S. Schweter. BERTurk: BERT models for Turkish. *Zenodo*, 2020.

[9] S. Yıldırım. Turkish-base-bert-sentiment-cased. *HuggingFace Model Hub*, 2020.

[10] YTÜ-CE Cosmos. Turkish-base-bert-uncased. *HuggingFace Model Hub*, 2023.

[11] VNGRS-AI. Kumru-2B: A Turkish Decoder Model. *HuggingFace Model Hub*, 2025.

[12] P.-J. Kindermans, et al. The (un)reliability of saliency methods. *Explainable AI*, Springer, 2019.

[13] N. Elhage, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[14] K. Wang, et al. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *ICLR*, 2023.

[15] K. Bostrom and G. Durrett. Byte Pair Encoding is Suboptimal for Language Model Pretraining. *EMNLP*, 2020.

[16] L. Xue, et al. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *TACL*, 2022.

[17] J. Vig, et al. Investigating gender bias in BERT's attention heads. *NeurIPS*, 2020.

[18] A. Geiger, et al. Causal abstractions of neural networks. *NeurIPS*, 2021.

[19] A. A. Akın and M. D. Akın. Zemberek, an open source nlp framework for turkic languages. *Structure*, 2007.

[20] A. Ustun, et al. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. *arXiv*, 2024.

[21] A. Vaswani, et al. Attention is all you need. *NeurIPS*, 2017.

[22] J. Devlin, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[23] J. DeYoung, et al. Eraser: A benchmark to evaluate rationales and explanations in nlp. *ACL*, 2020.

[24] A. Madsen, et al. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, 2021.

A Appendix: Mathematical Proofs

A.1 Proof of Theorem 1

Theorem 2 (Preservation of Completeness)**.** *Let ϕ_{tok} be a token-level attribution satisfying $\sum \phi_{tok} = \Delta F$. If the Morphological Alignment Matrix $\mathbf{A}$ satisfies the Strict Partition Property, then the projected attribution $\phi_{morph} = \mathbf{A}\phi_{tok}$ also satisfies Completeness.*

Proof. The proof relies on the linearity of the projection. However, a critical requirement is the handling of non-morpheme tokens (e.g., `[CLS]`, `[SEP]` in BERT-like models).

Let $\mathcal{T} = \{t_1, ..., t_T\}$ be the set of input tokens. We partition $\mathcal{T}$ into morphemic tokens $\mathcal{T}_m$ and special structural tokens $\mathcal{T}_s$. We construct $\mathbf{A}$ such that:

1. For $t_j \in \mathcal{T}_m$, $A_{kj} = 1$ iff t_j is part of morpheme μ_k.
2. For $t_j \in \mathcal{T}_s$, $A_{jj} = 1$ (Identity mapping), treating special tokens as atomic units.

Under this construction, the column-sum property $\sum_{k=1}^{K} A_{kj} = 1$ holds for all $j \in \{1, ..., T\}$. Therefore:

$$\sum_{k=1}^{K} \phi_{morph}^{(k)} = \sum_{k=1}^{K} \sum_{j=1}^{T} A_{kj} \phi_{tok}^{(j)} \tag{5}$$

$$= \sum_{j=1}^{T} \phi_{tok}^{(j)} \underbrace{\left(\sum_{k=1}^{K} A_{kj} \right)}_{=1 \text{ (Partition Property)}} \tag{6}$$

$$= \sum_{j=1}^{T} \phi_{tok}^{(j)} = F(x) - F(x') \tag{7}$$

Thus, completeness is preserved across the projection from computational to linguistic basis. □

B Appendix: Additional Qualitative Examples

Table 3 provides further examples comparing token-level IG with MAFEX across different linguistic phenomena.

Table 3: Additional Qualitative Comparisons.

Phenomenon	Input & Explanation
Derivation	*Gözlükçü* (Optician)
IG	Focuses on *Göz* (Eye)
MAFEX	Focuses on *-çü* (Occupation marker)
Double Neg.	*Yapmamış değilim* (I didn't not do it)
IG	Scattered across *yap*, *ma*, *değil*
MAFEX	Highlights both *-ma* and *değil* correctly.

C Appendix: Hyperparameter Sensitivity

We analyzed the impact of λ in Eq. 5. A value of $\lambda = 1.0$ (pure gradient) yields high sensitivity but low faithfulness. $\lambda = 0.0$ (pure causal) is faithful but ignores model internal mechanics. We found $\lambda = 0.7$ to be the optimal trade-off for Turkish morphology.

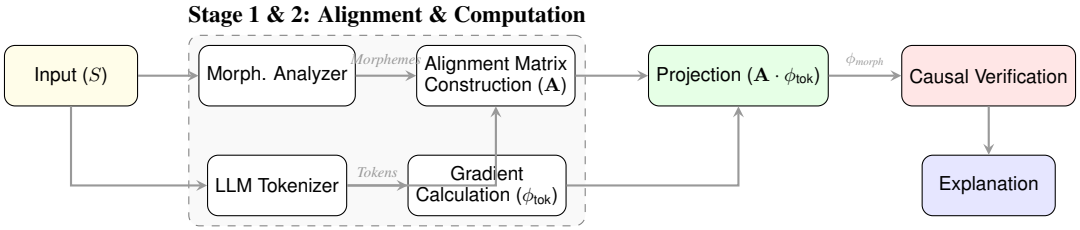

Figure 4: **System Pipeline.** The Alignment Matrix $\mathbf{A}$ bridges the gap between linguistic analysis (Zemberek) and neural computation (Tokens), enabling faithful projection before causal filtering.