PhO-Compress: A Two-Stage Framework to Enhance Optical LLM Context Compression

Christian Heinrich Hohlfeld
Konstanz, Germany, Born April 5, 1983
Bachelor of Science in Software Engineering
ORCID: https://orcid.org/0009-0003-6634-9045

October 25, 2025 (v2 Revision)

Abstract

Optical context compression, as pioneered by DeepSeek-OCR, presents a powerful but lossy method for scaling LLMs. This paper argues from first principles that the rate-distortion performance of such single-stage optical encoders is fundamentally limited by the high linguistic redundancy of their raw text input. We introduce PhO-Compress, a deterministic, two-stage framework invented by Christian Heinrich Hohlfeld, which acts as a lossless pre-optimizer to solve this problem. Stage I (Linguistic Reduction) decouples the input text into a normalized phonetic (IPA) stream and a lossless side-channel U. This side-channel perfectly preserves all structural, orthographic, and homophonic information, guaranteeing an orthographically-lossless reconstruction ($D_L = 0$). A "PhenomEncoder" then applies reversible predictive coding to the phonetic stream, achieving a lossless linguistic compression $\mathcal{C}_L$. Stage II (Optical Compression) renders only the reduced phonetic sequence Z_{Phenom} and applies a standard lossy visual encoder ($\mathcal{E}_V$), incurring distortion D_V. The total compression $\mathcal{C}_{Total} = \mathcal{C}_L \cdot \mathcal{C}_V$ is multiplicative. By removing linguistic redundancy *before* the optical stage, PhO-Compress enables $\mathcal{E}_V$ to achieve a superior rate-distortion trade-off, either reaching a higher compression $\mathcal{C}_V$ for the same D_V or a lower D_V for the same $\mathcal{C}_V$.

1 Introduction

The scaling of Large Language Models (LLMs) is fundamentally constrained by the quadratic complexity $\mathcal{O}(L^2)$ of the self-attention mechanism (Vaswani et al., 2017), despite optimizations like FlashAttention (Dao et al., 2023). A promising and potent solution is *optical context compression*, employed by models like DeepSeek-OCR (Wei et al., 2025) and Nougat (Blecher et al., 2023). These methods render text into compact images to be processed by an efficient visual backbone, effectively bypassing the $\mathcal{O}(L^2)$ text-processing bottleneck.

This paper argues from first principles that these current single-stage encoders are information-theoretically sub-optimal. They apply a single lossy compression function ($\mathcal{E}_V$) to a highly redundant input: raw text. This forces the encoder to simultaneously compress two distinct types of redundancy: (1) low-entropy *visual* redundancy from the text rendering, and (2) high-entropy *linguistic* redundancy (e.g., syntax, grammar, semantic predictability). We

hypothesize that this conflation leads to an inefficient rate-distortion trade-off; the encoder must expend ”bits” (or model capacity) to compress linguistic patterns that could be removed more efficiently.

We propose **PhO-Compress**, a two-stage framework that decouples these redundancies. It functions as a lossless, information-preserving pre-optimizer for any optical compression system.

The Grapheme-to-Phoneme (G2P) transformation is the critical first step that enables this decoupling. It provides three key advantages: (1) It separates semantics from orthography (case, punctuation), allowing orthography to be compressed losslessly in a separate side-channel U. (2) It maps homophones (e.g., ”there”, ”their”) to a single phonetic representation, which is perfectly disambiguated by a small bit-index in U. (3) It normalizes the text into a uniform alphabet (IPA) (International Phonetic Association, 1999), creating a statistically simpler stream for subsequent compression.

The two-stage pipeline is thus:

1. **Stage I (Orthographically-Lossless Reduction $\mathcal{E}_L$):** We apply a G2P transform $\mathcal{T}_{G2P_U}$ to get a phonetic stream X_{IPA_norm} and a lossless side-channel U. A reversible ”PhenomEncoder” $\mathcal{E}_L$ (see Sec 2.4) then removes predictable phonetic information (e.g., based on Zipf (1949)), yielding a compressed stream Z_{Phenom}. This stage is fully reversible and incurs zero distortion ($D_L = 0$).
2. **Stage II (Lossy Optical Compression $\mathcal{E}_V$):** The resulting sparse ”Phenom” sequence Z_{Phenom} is rendered and fed to a standard optical encoder (e.g., DeepSeek-OCR’s backbone). This stage incurs the system’s *only* distortion, D_V.

Our hypothesis is that by feeding the optical encoder $\mathcal{E}_V$ a pre-compressed, linguistically dense sequence Z_{Phenom}, $\mathcal{E}_V$ can focus *only* on the visual compression task it was designed for. This allows it to achieve a much higher effective compression ratio for a given distortion budget D_V than it could on the original, redundant L_{Text}.

2 Architecture of PhO-Compress

The pipeline is a deterministic chain of one lossless and one lossy encoder, preceded by necessary transformations.

$$X_{Text} \xrightarrow{\mathcal{T}_{G2P_U}} (X_{IPA\_norm}, U) \xrightarrow{\mathcal{E}_L} (Z_{Phenom}, U) \xrightarrow{\mathcal{R}} (I_{Phenom}, U) \xrightarrow{\mathcal{E}_V} (Z_{Vision}, U)$$

Note: The lossless side-channel U is extracted at the start and carried in parallel.

2.1 Stage I: Orthographically-Lossless Linguistic Reduction ($\mathcal{E}_L$)

The input X_{Text} is first processed by a deterministic and fully reversible transformation $\mathcal{T}_{G2P_U}$. This function decouples the text into two distinct streams: a normalized phonetic input X_{IPA_norm} and a comprehensive, lossless side-channel U.

$$(X_{IPA\_norm}, U) = \mathcal{T}_{G2P_U}(X_{Text})$$

The side-channel U captures all information lost during the normalization and G2P process. Its rigorous definition is provided in Section 3.3.

The PhenomEncoder $\mathcal{E}_L$ (the "linguistic reduction") then operates *only* on the normalized phonetic stream X_{IPA_norm}:

$$Z_{Phenom} = \mathcal{E}_L(X_{IPA\_norm})$$

Definition 2.1 (Orthographically Lossless Reconstruction ($D_L = 0$))**.** *The $D_L = 0$ mode is the primary operational mode of this framework, guaranteeing bit-exact reconstruction of the original text. This is achieved by the combined decoders $(\mathcal{D}_L, \mathcal{T}_U^{-1})$.*

Let $\mathcal{D}_L$ be the deterministic linguistic decoder inverse to $\mathcal{E}_L$, such that $\hat{X}_{IPA\_norm} = \mathcal{D}_L(Z_{Phenom})$. Let $\mathcal{T}_U^{-1}$ be the deterministic inverse transformation (re-assembler) that uses the side-channel U.

The total reconstruction is $\hat{X}_{Text} = \mathcal{T}_U^{-1}(\hat{X}_{IPA\_norm}, U)$.

The framework is orthographically lossless ($D_L = 0$) because $\mathcal{E}_L$ and $\mathcal{D}_L$ are designed as a fully reversible pair ($\hat{X}_{IPA\_norm} \equiv X_{IPA\_norm}$) AND the side-channel U is comprehensive, guaranteeing:

$$\hat{X}_{Text} = \mathcal{T}_U^{-1}(\mathcal{D}_L(\mathcal{E}_L(\mathcal{T}_{G2P_U}(X_{Text})_1)), \mathcal{T}_{G2P_U}(X_{Text})_2) \equiv X_{Text}$$

In this mode, $\mathcal{E}_L$ only removes phonetic/linguistic redundancies that are perfectly predictable and reversible by $\mathcal{D}_L$, while U preserves all orthographic and structural uniqueness.

Remark 2.2 (Semantic-Lossy ($D_L > 0$) Variant)**.** *A secondary, semantic-lossy mode is theoretically possible. In this variant, $\mathcal{E}_L$ would be a lossy encoder (e.g., a "semantic summarizer" in the phonetic domain) that discards information to satisfy a semantic distortion budget $D_L > 0$. However, this requires a correspondingly complex decoder $\mathcal{D}_L$ (likely a large LM itself) to "in-fill" the lost meaning, making it computationally expensive and difficult to control. We therefore focus on the $D_L = 0$ mode.*

2.2 Stage II: Lossy Optical Compression ($\mathcal{E}_V$)

The reduced symbolic sequence Z_{Phenom} is rendered by a deterministic, lossless function $R : Z_{Phenom} \to \mathcal{I}_{Phenom}$ into a compact image representation I_{Phenom}. This image is then encoded by a lossy visual encoder $\mathcal{E}_V$ (e.g., a ViT or VLM backbone).

$$Z_{Vision} = \mathcal{E}_V(I_{Phenom}) = \mathcal{E}_V(\mathcal{R}(Z_{Phenom}))$$

This stage incurs a visual reconstruction distortion D_V.

Remark 2.3 (Fine-Tuning Requirement)**.** *The visual encoder $\mathcal{E}_V$ and its corresponding decoder $\mathcal{D}_V$ (e.g., an OCR model) are typically pre-trained on images of natural text (graphemes). To achieve optimal performance and minimize D_V, these models would require fine-tuning on a synthetic dataset of rendered IPA sequences (I_{Phenom}) generated by $\mathcal{R}$.*

2.3 Total Distortion

In the primary $D_L = 0$ mode, the only source of loss is the optical stage.

Definition 2.4 (Total Distortion ($D_L = 0$ Mode)). *The total semantic distortion D_{Total} is the error introduced by the optical stage. Let the full decoder be $\hat{X} = \mathcal{T}_U^{-1}(\mathcal{D}_L(\mathcal{R}^{-1}(\mathcal{D}_V(Z_{Vision}))), U)$. Since $\mathcal{T}_U^{-1}$, $\mathcal{D}_L$, and $\mathcal{R}$ are lossless, the total distortion is:*

$$D_{Total} = \mathbb{E}[d_S(S(X_{Text}), S(\hat{X}))] \leq D_V$$

Where D_V is the semantic impact of visual loss from $\mathcal{E}_V$. The goal is to choose a D_V that is sufficiently low for the target task S.

2.4 Proposed Architecture for the $\mathcal{E}_L/\mathcal{D}_L$ Pair

The primary challenge of Stage I is to design a reversible pair $(\mathcal{E}_L, \mathcal{D}_L)$ that achieves non-trivial compression ($\mathcal{C}_L > 1$). We propose a concrete architecture based on predictive coding.

Definition 2.5 (Predictive PhenomEncoder). *We can model $\mathcal{E}_L$ and $\mathcal{D}_L$ as identical, deterministic, predictive models (e.g., a simple Markov model, n-gram model, or a small, fixed-weight RNN) trained on the statistics of the phonetic stream X_{IPA_norm}.*

1. ***Encoder ($\mathcal{E}_L$):*** *The encoder iterates through the input X_{IPA_norm}. At each step t, it uses its internal state (based on $p_{t-1}, p_{t-2}, ...$) to predict the next phoneme $\hat{p}_t$.*
    - *If the prediction is correct (i.e., $\hat{p}_t == p_t$), the encoder outputs nothing (or a 1-bit "match" token) to Z_{Phenom}.*
    - *If the prediction is incorrect, the encoder outputs the *actual* phoneme p_t (or a "mismatch" token followed by p_t) to Z_{Phenom}.*

    *It then updates its state with the *actual* phoneme p_t to remain synchronized with the decoder.*

2. ***Decoder ($\mathcal{D}_L$):*** *The decoder is an identical model. It reads the compressed stream Z_{Phenom}.*
    - *If it reads a "match" token, it knows its prediction $\hat{p}_t$ was correct, so it outputs $\hat{p}_t$ to the reconstructed stream $\hat{X}_{IPA\_norm}$.*
    - *If it reads a "mismatch" token (or a full phoneme), it knows its prediction was wrong, so it reads the *actual* phoneme p_t from the stream and outputs p_t.*

    It then updates its state with p_t, remaining perfectly synchronized with the encoder.

This architecture is fully and deterministically reversible ($D_L = 0$) and achieves compression $\mathcal{C}_L$ proportional to the predictability of the phonetic stream.

3 Mathematical Foundations

The information-theoretic bounds govern the *entire* system $E = \mathcal{E}_V \circ \mathcal{R} \circ \mathcal{E}_L \circ \mathcal{T}_{G2P_U}$.

3.1 Impossibility of Universal Lossless Compression

Theorem 3.1 (No Universal Lossless Compression)**.** *There exists no lossless code C that can guarantee a compression rate $r < 1$ for all possible input strings x (Kolmogorov (1965); Cover and Thomas (2006)).*

Implication: This theorem is why PhO-Compress must be a lossy framework ($D_V > 0$) to guarantee compression ($\mathcal{C}_{Total} > 1$) on all inputs. The $D_L = 0$ stage only re-arranges information and removes predictable redundancy.

3.2 Rate-Distortion Bounds (End-to-End)

The theoretical limit for the *entire* PhO-Compress pipeline is the semantic rate-distortion function for the task S.

Theorem 3.2 (Zero-Distortion Lower Bound)**.** *If the task S must be preserved perfectly ($D_{Total} = 0$), the minimum achievable bit-rate R is bounded by the entropy of the task representation:*

$$R = \frac{1}{n}\mathbb{E}|Z_{Vision}| \geq \frac{1}{n}H(S(X^n))$$

This requires both $\mathcal{E}_L$ and $\mathcal{E}_V$ to be S-lossless (Shannon, 1948).

Theorem 3.3 (Semantic Rate-Distortion)**.** *For a tolerable distortion $D_{Total} > 0$ (where $D_{Total} \leq D_V$), the optimal asymptotic rate R for the end-to-end system is given by the semantic rate-distortion function (?):*

$$R_S(D_{Total}) = \inf_{P_{\hat{X}|X}:\mathbb{E}[d_S(S(X),S(\hat{X}))]\leq D_{Total}} I(S(X);\hat{X})$$

PhO-Compress attempts to approximate this bound by optimizing $\mathcal{C}_L$ (which is lossless, $D_L = 0$) and $\mathcal{C}_V$ (which is lossy, $D_V > 0$).

3.3 Side-Channel U for Lossless Reconstruction

The lossless side-channel U is the key to the $D_L = 0$ mode. We provide a rigorous definition.

Definition 3.4 (Comprehensive Side-Channel U)**.** *The side-channel U is a deterministic, lossless data stream generated by the parser $\mathcal{T}_{G2P_U}$. It captures all information required to invert the phonetic normalization and reduction, ensuring $\hat{X}_{Text} \equiv X_{Text}$. U is a composite of disjoint streams: $U = (U_{struct}, U_{ortho})$.*

1. U_{struct} ***(Structural/High-Frequency Data):*** *This stream losslessly encodes structure-critical tokens such as code, numbers, URLs, and other spans ill-suited for phonetic processing.*
2. U_{ortho} ***(Orthographic Disambiguation):*** *This stream captures all orthographic and layout information removed during normalization.*
    - U_{case}***:*** *A bit-mask (e.g., 2 bits/token) storing the capitalization state (lower, Title, UPPER, MiXeD) for all tokens.*

- $U_{punct/ws}$***: A deterministic representation (e.g., RLE-encoded) of all inter-token whitespace and punctuation.***
- U_{hom} ***(Homophone/Homograph Disambiguation):*** *For any phonetic sequence p generated by $\mathcal{T}_{G2P_U}$ that maps to a homophone class $H(p) = \{w_1, \ldots, w_k\}$ with $k > 1$, U_{hom} stores a unique index $i \in [0..k-1]$ for the original grapheme w_i. The bit-cost for this token is $\lceil \log_2 k \rceil$. (e.g., for* `(there, their, they're)`*, $k = 3$, cost=2 bits).*
- U_{oov}***: A dictionary for any Out-Of-Vocabulary graphemes that the G2P model $\mathcal{T}_{G2P_U}$ cannot handle deterministically.***

Information-Theoretic Implication: The final representation of Stage I is the joint pair (Z_{Phenom}, U). The minimum rate R for $D_{Total} = 0$ (assuming $D_V = 0$ as well) is bounded by the joint entropy of the *full* information:

$$R = \frac{1}{n}\mathbb{E}|(Z_{Vision}, U)| \geq \frac{1}{n}H(X^n)$$

The goal of PhO-Compress is to minimize the size of Z_{Phenom} via $\mathcal{E}_L$ (which is passed to the lossy $\mathcal{E}_V$) and leverage the fact that $H(U) \ll H(X^n)$ (i.e., the orthographic side-channel is information-theoretically small and can be passed losslessly).

4 Proportionality and Multiplicative Compression

We define the compression factors based on symbolic lengths (e.g., tokens, phonemes, visual patches). Let L_{Text} be the number of text tokens, L_{Phenom} the number of phenom-symbols, and L_{Vision} the number of final visual tokens.

Definition 4.1 (Compression Factors)**.** *The Stage I (Linguistic) compression factor is:*

$$\mathcal{C}_L = \frac{L_{Text}}{L_{Phenom}}$$

The Stage II (Optical) compression factor is:

$$\mathcal{C}_V = \frac{L_{Phenom}}{L_{Vision}}$$

Proposition 4.2 (Multiplicative Compression)**.** *The total symbolic compression factor is the product of the stage-wise factors:*

$$\mathcal{C}_{Total} = \frac{L_{Text}}{L_{Vision}} = \frac{L_{Text}}{L_{Phenom}} \cdot \frac{L_{Phenom}}{L_{Vision}} = \mathcal{C}_L \cdot \mathcal{C}_V$$

Remark 4.3 (Non-Triviality)**.** *This multiplication is non-trivial. It posits that total compression is the product of two distinct reduction strategies. The core hypothesis of PhO-Compress is that $\mathcal{C}_{Total}$ can be maximized by optimizing $\mathcal{C}_L$ and $\mathcal{C}_V$ jointly. By feeding $\mathcal{E}_V$ a linguistically sparse input Z_{Phenom}, we hypothesize $\mathcal{E}_V$ can achieve a higher effective $\mathcal{C}_V$ (or lower D_V) than it could by directly processing the redundant X_{Text}.*

5 Complexity and Decoder Cost Analysis

The viability of this framework hinges on the decoder cost being substantially lower than the savings from attention.

Definition 5.1 (Revised Total Cost)**.** *The total inference cost T_{tot} includes all encoding, decoding, and final LLM processing. The decoders are split into the linguistic/phonetic decoder $\mathcal{D}_L$ and the deterministic orthographic re-assembler $\mathcal{T}_U^{-1}$.*

$$T_{tot} = T(\mathcal{T}_{G2P_U}) + T(\mathcal{E}_L) + T(\mathcal{R}) + T(\mathcal{E}_V) + T(\mathcal{D}_V) + T(\mathcal{D}_L) + T(\mathcal{T}_U^{-1}) + T_{LLM}(L_{Vision})$$

Where $T(\mathcal{T}_{G2P_U})$ is the G2P and U extraction, $T(\mathcal{D}_V)$ is the optical decoder/OCR, $T(\mathcal{D}_L)$ is the linguistic reconstructor, and $T(\mathcal{T}_U^{-1})$ is the re-injection of the side-channel U.

Hypothesis 5.2 (Practical Viability in $D_L = 0$ Mode)**.** *The PhO-Compress framework is viable if the pre- and post-processing overhead is less than the savings from $\mathcal{O}(L^2)$ attention:*

$$T(\mathcal{T}_{G2P_U}, \mathcal{E}_L, ..., \mathcal{D}_L, \mathcal{T}_U^{-1}) < T_{LLM}(L_{Text}) - T_{LLM}(L_{Vision})$$

This hypothesis is strong in the $D_L = 0$ mode. Here, both $T(\mathcal{D}_L)$ (the predictive decoder from Sec 2.4) and $T(\mathcal{T}_U^{-1})$ (dictionary lookups, string appends) are computationally cheap, low-complexity operations. The viability cost is dominated by the optical $T(\mathcal{D}_V)$ (OCR), which is a known, manageable cost, rather than a large, semantic LM.

6 Positioning and Related Work

Enhancement for Optical Context Compression. PhO-Compress is not a competitor to optical methods like Wei et al. (2025) or Blecher et al. (2023); it is a **pre-optimizer** for them. These systems define the $X_{Text} \to \mathcal{E}_V \to Z_{Vision}$ pipeline. Our framework enhances this pipeline by transforming it into $X_{Text} \to (\mathcal{E}_L, U) \to \mathcal{E}_V \to Z_{Vision}$. We provide a method to losslessly 'clean' and compress the input *before* it hits their encoder, thereby improving the $\mathcal{C}_V$ they can achieve for any given D_V.

Prompt Compression. Jiang et al. (2LOS 2023) and Jiang et al. (2024) perform a function analogous to our $\mathcal{E}_L$ (S-sufficient reduction) but operate directly in the text-token space, not the phonetic space, and do not typically guarantee orthographically-lossless reconstruction as our U-channel does.

7 Conclusion

The PhO-Compress framework, invented by Christian Heinrich Hohlfeld, provides a novel, first-principles path to enhancing optical context compression. By decoupling the problem into two distinct stages—(I) reversible phonetic reduction $(\mathcal{E}_L)$ and (II) lossy optical encoding $(\mathcal{E}_V)$—it solves the information-theoretic bottleneck of single-stage encoders.

The primary innovation is the $D_L = 0$ orthographically-lossless Stage I, which functions as a pre-optimizer. It uses a lossless side-channel U to solve the ambiguity and information-loss problems that plague purely semantic approaches. We posit that this two-stage ap-

proach is not merely an alternative, but a necessary step to achieve the theoretical rate-distortion limits for optical compression by allowing each stage to specialize: $\mathcal{E}_L$ on linguistic redundancy and $\mathcal{E}_V$ on visual redundancy.

Availability & Licensing

A reference implementation will be released under a dual license (AGPL-3.0 + Commercial); model weights under a responsible-AI license; datasets under ODC-BY 1.0. For enterprise licensing and support, please contact the author.

License & Use

Acknowledgments

I express heartfelt gratitude to my parents, Barbara Friderike Hohlfeld and Dr. med. Siegfried Hohlfeld, for their unwavering love, support, and encouragement throughout my life. Their belief in my potential has been instrumental in the pursuit of knowledge, innovation, and ideas. The author also acknowledges the contributions of the Gemini (Google) and ChatGPT (OpenAI) AI systems for assistance with collaborative refinement, structural verification, code-generation, proof generation, and proofreading of the author's original work.

References

Blecher, Lukas, Guillem Cucurull, Thomas Scialom, and Robert Stojnic (2023). "Nougat: Neural Optical Understanding for Academic Documents". In: *arXiv: 2308.13418*.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of Information Theory*. 2nd ed. Wiley.

Dao, Tri, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré (2023). "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness". In: *ICML*. arXiv: 2205.14135.

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Jiang, Hui et al. (2024). "LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression". In: *ACL 2024 (Long Papers)*. DOI: 10.18653/v1/2024.acl-long.91.

Jiang, Huiqiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu (2023). "LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models". In: *arXiv: 2310.05736*.

Kolmogorov, Andrey N. (1965). "Three Approaches to the Quantitative Definition of Information". In: *Problemy Peredachi Informatsii* 1.1, pp. 3–11.

Shannon, Claude E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27, pp. 379–423, 623–656. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Shannon, Claude E. (1959). "Coding Theorems for a Discrete Source with a Fidelity Criterion". In: *IRE National Convention Record*. Vol. 7, pp. 142–163.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention Is All You Need". In: *NeurIPS*. arXiv: 1706.03762.

Wei, Haoran, Chenglong Liu, et al. (2024). "General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model". In: *arXiv: 2409.01704*.

Wei, Haoran, Yaofeng Sun, Yukun Li, et al. (2025). "DeepSeek-OCR: Contexts Optical Compression". In: *Initial investigation into visual long-context compression*. arXiv: 2510.18234.

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

A Appendix: Didactic Stage-I Prototype (Python)

The following code provides a simplified, didactic prototype of the Stage I PhenomEncoder ($\mathcal{E}_L$). It demonstrates the $X_{Text} \to X_{IPA} \to Z_{Phenom}$ chain by using simple stop-word removal as a proxy for true semantic reduction. (Note: This prototype omits the rigorous side-channel U implementation for clarity).

```python
import re
from typing import List, Dict

class PhenomEncoderPrototype:
    """Simulates the linguistic compression (E_L) via simple filters."""

    def __init__(self, ipa_map: Dict[str, str], stop_words: List[str]):
        self.ipa_map = ipa_map
        self.stop_words = set(stop_words)
        # This regex is a placeholder for a real U_struct extractor
        self.critical_pattern = re.compile(r'\b\d+[\d,.]*\b|[A-Z_]{2,}')

    def _g2p_transform(self, text: str) -> (List[str], List[str]):
        """Didactic G2P and word tokenizer."""
        words = re.findall(r'\b\w+\b', text.lower())
        ipa_tokens = []
        for w in words:
            ipa_tokens.append(self.ipa_map.get(w, f'<{w}>'))
        return ipa_tokens, words

    def encode_linguistic(self, text: str) -> List[str]:
        """
        Performs Stage I (E_L) reduction.
        This example uses stop-word removal as a simple proxy
        for S-sufficient reduction.
        """

        ipa_tokens, original_words = self._g2p_transform(text)

        phenom_tokens = []

        for word, ipa in zip(original_words, ipa_tokens):
            # A real implementation would check U_struct here
            # This proxy filter removes stop words and OOV tokens
            if word not in self.stop_words and not ipa.startswith('<'):
                phenom_tokens.append(ipa)

        return phenom_tokens
```

```python

if __name__ == '__main__':
    # Didactic dictionary (Example IPA)
    IPA_DICTIONARY = {
        "die": "/di/", "der": "/de?r/", "ist": "/?st/", "wurde": "/vUrd@/",
        "system": "/zYs'te:m/", "architektur": "/?arCitek'tu:r/",
        "neue": "/'nOY@/", "hybrid": "/hy'bri:t/", "entworfen": "/?Ent'vOrf@n/"
    }

    STOP_WORDS_LIST = ["die", "der", "ist", "und", "ein", "eine",
                       "das", "als", "wurde"]

    enc = PhenomEncoderPrototype(IPA_DICTIONARY, STOP_WORDS_LIST)

    text = "Die neue Hybrid Architektur wurde als ein System entworfen."

    z_phenom = enc.encode_linguistic(text)

    orig_len_words = len(text.split())
    phenom_len_tokens = len(z_phenom)

    c_l = (orig_len_words / phenom_len_tokens) if phenom_len_tokens > 0 else 0

    print(f"Original Text (Words): {orig_len_words}")
    print(f"ZPhenom (Tokens): {phenom_len_tokens}")
    print(f"ZPhenom Stream: {' '.join(z_phenom)}")
    print(f"C_L (Didactic Semantic Rate): {c_l:.2f}x")
    print("\nThis ZPhenom sequence would now go to Stage II (Rendering")
    print("and lossy optical encoding).")
```

Listing 1: Simplified PhenomEncoder (Stage I) Prototype