ATTENTION'S GRAVITATIONAL FIELD: A POWER-LAW AND LINGUISTIC INTERPRETATION OF POSITIONAL CORRELATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Positional encodings are crucial for Transformer architectures, yet their underlying geometric and linguistic principles remain poorly understood. In this work, we introduce the **Attention Gravitational Field (AGF)**, a principled relative positional framework that decouples geometric spatial relationships from semantic content using a power-law formulation. We demonstrate a striking mathematical convergence between linguistic distance decay, empirical attention behavior, and classical Newtonian physics—specifically the Universal Law of Gravitation and the Spherical Shell Theorem. To bridge continuous physical fields with discrete syntactic boundaries, we propose Single-Shell (**AGF-S**) and Multi-Shell (**AGF-MS**) extensions that resolve long-standing limitations such as the self-attention zero-distance paradox and directional asymmetry. Empirical evaluations across machine translation benchmarks confirm that AGF-MS yields consistent BLEU gains over standard sinusoidal encodings and prior relative decay methods (e.g., ALiBi and KERPLE). Ultimately, our work offers a unified physical and linguistic perspective on positional representation, opening new avenues for interpretable and physically grounded Transformer architectures.

1 INTRODUCTION

Understanding positional relationships lies at the very heart of the Attention mechanism and the Transformer architecture. From early absolute positional encodings (Vaswani et al., 2017) to subsequent relative variants (Shaw et al., 2018; Su et al., 2024), researchers have provided diverse theoretical and empirical foundations that now power modern Large Language Models (LLMs). However, to date, no existing work has critically analyzed or demonstrated the fundamental nature of attention positionality grounded in linguistic common sense.

This paper represents a novel, groundbreaking attempt to bridge that gap. By analyzing the deep semantics of attention alongside linguistic structure, we establish that the core mechanism is essentially a minimal-cost encoding model for word pairing and modification. Its intrinsic properties exhibit striking similarities to Zipf's Law, reliability growth models, and standard learning curves, making power-law decay an ideal representation. Furthermore, drawing an analogy to Newtonian gravitation, we derive a gravitational field formulation tailored for Attention, termed **AGF** (*Attention's Gravitational Field*). While mathematically equivalent to existing formulations such as KERPLE (Chi et al., 2022), AGF offers superior conceptual clarity, a rigorous step-by-step derivation, and a closer alignment with physical principles.

Extending this analogy to the Newtonian spherical shell model and incorporating linguistic features, we present a refined variant named **AGF-S**. Upon visualizing its behavior, we identified its functional limitations and further integrated linguistic common sense to derive a multi-shell variant, **AGF-MS**.

We evaluate these theoretical models within a Transformer framework across two benchmarks: English–German (En–De) and English–Romanian (En–Ro). The empirical results strongly support our theoretical conjectures, achieving a 0.3–0.5 BLEU score improvement over strong baselines including ALiBi (Press et al., 2022) and log@KERPLE. Beyond these numerical gains, our cen-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tral contribution is the establishment of a coherent, highly interpretable theoretical framework that faithfully aligns with fundamental linguistic structure and syntax rules.

2 Related Work

In early Transformer architectures and Large Language Models (LLMs) (Vaswani et al., 2017), positional information was incorporated by adding absolute positional encodings directly into semantic embeddings. Conceptually, element-wise addition of position and semantics is fundamentally ungrounded—much like adding *"age"* to *"income"*—causing severe semantic distortion and representation confusion.

Subsequent innovations sought to isolate or constrain this positional-semantic interference. For instance, Rotary Position Embedding (RoPE) (Su et al., 2024) confines the interaction between position and semantics to specific projection stages within the complex domain. To enforce spatial attenuation, ALiBi (Press et al., 2022) introduces a static exponential decay penalty, whereas the T5 architecture (Raffel et al., 2020) utilizes bucketed, piecewise distance parameters to fit relative positional relationships. Taking a function-space perspective, KERPLE (Chi et al., 2022) leverages kernelized formulations to offer a broader suite of candidate decay curves.

While these foundational works have demonstrated compelling empirical performance, a pervasive limitation persists: they rarely incorporate linguistic common sense, such as part-of-speech (POS) tags or syntactic dependencies. To date, no open literature has critically explored or rigorously demonstrated the underlying linguistic nature governing positional relationships in attention.

3 Attention's Gravitational Field

3.1 What is Attention?

In natural language, conventional linguistic analysis categorizes words into part-of-speech (POS) tags—such as nouns, verbs, adjectives, adverbs, prepositions, and quantifiers—which are further organized into fine-grained syntactic structures (e.g., subject–verb, subject–verb–object, and passive constructions). Across these structures, functional pairings and modification rules remain remarkably consistent: adjectives modify nouns, adverbs modify verbs, degree adverbs modify adjectives, and nouns preceding a verb predominantly function as subjects (see Figure 1). Fundamentally, syntactic parsing is the search for underlying structural regularities. These "regularities"—co-occurrence, linkage, modification, and coordination patterns that appear with high probability in human discourse—constitute the true linguistic essence of Attention.

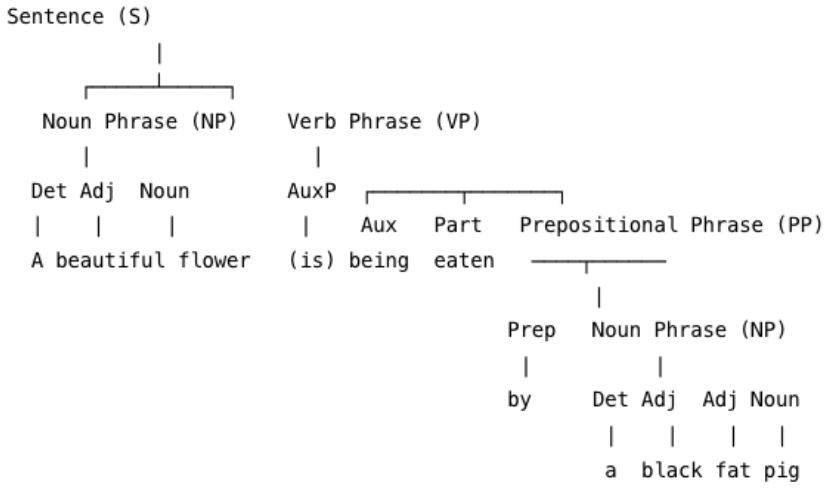

Figure 1: Syntactic dependency and part-of-speech (POS) modification patterns illustrating the spatial distance dynamics between modifiers and target heads in natural language.

Furthermore, fine-grained semantic subcategories reveal sharper probabilistic dynamics. Consider a scenario where *"beautiful"* serves as the Query (Q) and animal/plant nouns serve as the Value (V). In massive corpora, the conditional probability distribution of the immediate next token (pos $= 1$) yields an attention score distribution $P(\text{pos} = 1)$ with a cumulative density of 1. Here, $P(\text{girl} \mid \text{beautiful})$ significantly exceeds $P(\text{pig} \mid \text{beautiful})$.

However, as the relative distance expands to pos $= 9$, the contextual scope shifts:

"a (beautiful) x x x x x x x x (?)"

Consider a real-world corpus instantiation:

"a beautiful flower is being eaten by a black fat pig"

At pos $= 9$, the conditional distribution $P(\text{pos} = 9)$ may assign a higher score to *"pig"* than to *"girl"*. Because structural modification constraints attenuate over distance, out-of-scope token occurrences increase rapidly. Since the cumulative probability density remains bounded by 1, the target attention score exhibits a swift decay.

Conversely, when modifying relations span longer distances within a valid syntactic frame:

"a (beautiful) and elegant (girl) is singing"
"a (beautiful), but tired (girl) is working"

Intervening tokens refine semantic details without breaking the core modifier-head link between *"beautiful"* and *"girl"*. In human language, each token or attention pairing compresses an event by capturing its most frequent informational core. Additional modifiers are introduced selectively to refine details—analogous to increasing image resolution to reveal fine-grained features. For instance, *"An apple"* implicitly denotes a typical red, medium-sized fruit. Introducing modifiers like *"a green apple"* or *"a large green apple"* consumes additional sequence length to specify less frequent, detail-rich scenarios.

When aggregated across massive datasets, any attention or syntactic structure yields a decaying probability density curve over sequence length. We define this metric as the **Probability of Attention's Sequence Length (PASL)**, formulated as:

$$P_{\text{ASL}}(i), \quad \text{for } i \geq 0, \tag{1}$$

where i denotes the length expansion $(2 + i)$ for a given attention pattern.

Rather than exhibiting arbitrary noise, $P_{\text{ASL}}(i)$ demonstrates a systematic, monotonic decay. Grounded in Shannon entropy (Shannon, 1948) and Huffman coding principles, high-frequency semantics are encoded via minimal sequence lengths to achieve optimal communication efficiency:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \tag{2}$$

Human discourse inherently operates under the principle of linguistic economy. Speakers default to concise formulations for high-probability events, opting for extended modification structures only when lower-probability, detail-heavy descriptions become necessary.

3.2 Power-Law Dynamics in PASL

Zipf's Law and Linguistic Economy As early as Zipf's foundational statistical observations (Zipf, 1935), word frequencies across natural language corpora have been recognized to follow a power-law distribution governed by the principle of least effort. While Zipf's Law addresses global, corpus-wide lexical distributions, our proposed **PASL** models localized, structure-specific event descriptions. Despite operating at distinct granularities, both mechanisms share an identical foundation: the minimization of communication and encoding costs. Consequently, it is theoretically grounded to model the spatial decay of PASL via a power-law distribution.

Reliability Growth and Duane Modeling In reliability engineering, Duane (1964) demonstrated that cumulative failure rates decay linearly in log-log space under continuous system optimization. The continuous Duane model is traditionally formulated as:

$$\ln(\text{MTBF}) = -\ln a + m \ln t, \tag{3}$$

where MTBF denotes the Mean Time Between Failures, t represents cumulative operating time, and a, m are system parameters.

A structural parallel exists within the sequence length dynamics of Attention. The necessity for an extended context of length $i + 1$ arises only when a shorter context (i) fails to provide sufficient semantic precision—a state analogous to a system specification "failure." As complex, low-probability details are progressively resolved by additional modifiers, the likelihood of requiring further sequence expansion diminishes, directly mapping PASL onto reliability growth dynamics governed by power-law attenuation.

Learning Curves and Cumulative Density Functions Furthermore, PASL can be interpreted through the lens of classical learning curve theory. Each additional token slot acts to maximize information coverage over the remaining semantic variance, directly mirroring how successive training epochs optimize system parameters (see Figure 2). Mathematically, the cumulative distribution function (CDF) of PASL corresponds to the trajectory of a learning curve.

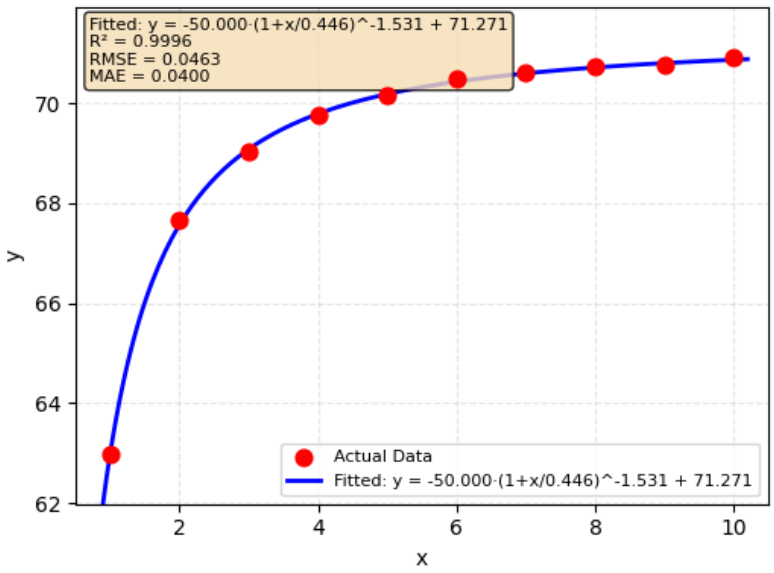

Figure 2: Trajectory of cumulative information coverage mirroring classical learning curves, illustrating how successive sequence expansions minimize semantic uncertainty under power-law dynamics.

Given its ubiquity across complex systems, a power-law formulation offers a principled foundation for relative positional decay in attention mechanisms.

3.3 Attention's Gravitational Field (AGF)

Derivation from Universal Gravitation As established in our analysis of PASL, every attention pair has a minimum token distance of $d = 1$ and obeys a fundamental power-law decay. To formalize this relationship, we draw an explicit analogy to Newtonian gravitation—one of the most ubiquitous power-law formulations in physics:

$$F = \frac{GMm}{r^2}. \tag{4}$$

For an object located at an elevation d above a spherical body of radius r, the gravitational attraction is expressed as:

$$F(M, m, d) = \frac{GMm}{(r+d)^2}. \tag{5}$$

Normalizing with respect to the baseline surface strength, the relative field attenuation ratio simplifies to:

$$F(d) = \text{Base} \cdot \frac{1}{(1+d/r)^2}. \tag{6}$$

Extending this formulation from three-dimensional physical space to the high-dimensional latent manifolds of transformer models requires parameterizing the spatial dimension exponent. This yields our **Attention's Gravitational Field (AGF)** formulation:

$$g(d) = \frac{1}{(1+d/r)^k}, \quad \text{for } k > 0, r > 0, d \geq 0, \tag{7}$$

where r operates as an intrinsic characteristic "radius" defining the structural capacity of the attention subspace.

Mathematical Equivalence to KERPLE Crucially, our physically derived AGF framework exhibits exact mathematical equivalence to empirical kernelized positional encodings, notably the log-variant of KERPLE (Chi et al., 2022):

$$k^{\text{comp}}([\boldsymbol{q}_m, m], [\boldsymbol{k}_n, n]) = \boldsymbol{q}_m^\top \boldsymbol{k}_n + c - r_1 \cdot \ln(1 + r_2|m-n|), \quad \text{with } r_1, r_2 > 0. \tag{8}$$

When mapped into the post-exponentiation domain prior to Softmax normalization, the structural isomorphism becomes transparent:

- The additive constant $\exp(c)$ maps to the gravitational baseline factor.
- The coefficient r_1 corresponds directly to the spatial dimension exponent k.
- The scaling parameter r_2 corresponds to the inverse characteristic radius $1/r$.

Thus, KERPLE's empirical formulation is structurally isomorphic to AGF. Throughout the rest of this paper, we maintain the multiplicative power-law notation $g(d)$ for theoretical grounding, converting into equivalent logarithmic additive biases during implementation to integrate seamlessly with the Softmax exponential.

4 Linguistic Analysis and Shell-Based Models

While our derived **AGF** framework establishes a formal equivalence between Universal Gravitation and logarithmic kernel decay (e.g., log@KERPLE (Chi et al., 2022)), analyzing attention dynamics through formal linguistic theory reveals crucial structural limitations in existing formulations.

4.1 Linguistic Property Analysis

Minimum Distance and the Spherical Shell Analogy In direct modifier-head constructions (*"beautiful girl"*), the minimal positional offset between Query and Key tokens is $d = 1$. However, in predicate constructions (*"she is beautiful"*), syntactic rules necessitate an intervening verb, establishing a mandatory minimum structural separation ($d \geq 2$).

This structural constraint exhibits a direct physical parallel to Newton's Spherical Shell Theorem: inside a uniform spherical shell, internal gravitational forces cancel to zero, whereas outside the shell, attraction follows standard power-law attenuation. Analogously, the inner shell radius represents the minimum token distance mandated by syntactic rules, inducing a discrete step-change in spatial attention affinity.

The Self-Attention Zero-Distance Paradox Existing relative positional encodings (Press et al., 2022; Chi et al., 2022) universally assume that token affinity reaches its global maximum at $d = 0$. From a linguistic perspective, this assumption is sub-optimal. A position of $d = 0$ represents self-attention (identity mapping). Over language evolution, frequently co-occurring modifier-head pairs naturally undergo **lexical unification** (*semantic compression*).

For instance, explicitly paired expressions requiring multi-token spatial offset—such as *"a beautiful person"* or *"a noble ruler"*—gradually condense into single, unified lexical units like *"beauty"* or *"king"*. These unified tokens inherently absorb the modifier's semantic attributes into a single representation. Consequently, while self-attention for such self-contained tokens may retain a non-zero baseline weight ($g_0 \in [0.3, 0.5]$), enforcing $d = 0$ as the absolute peak of the spatial attenuation curve conflates internal lexical unification with cross-token modifier-head interactions.

Directional Asymmetry in Relative Positioning Syntactic dependencies exhibit explicit directional asymmetry. Modifiers can function as pre-modifiers or post-modifiers (*"a beautiful girl"* vs. *"the girl is beautiful"*), possessing distinct structural distance distributions across left and right contexts. While causal Decoders only attend to past tokens ($d > 0$), bidirectional Encoders must parameterize left and right contexts independently to avoid unnatural symmetry constraints.

4.2 Single-Shell Model (AGF-S)

To capture these linguistic constraints, we parameterize the spatial decay profile. For unidirectional **Decoder** contexts, the piece-wise formulation is defined as:

$$g(d) = \begin{cases} g_0, & d = 0, \\ 0, & 0 < d < r_1, \\ \frac{1}{\left(1+\frac{d-r_1}{r_2}\right)^k}, & d \geq r_1, \end{cases} \tag{9}$$

where r_1 denotes the inner shell boundary, r_2 is the characteristic scaling radius, and $k > 0$ represents the spatial attenuation exponent.

For bidirectional **Encoder** contexts, we introduce directional scaling parameters (α_1, α_2) and independent shell radii (r_1, r_2 for left context; r_3, r_4 for right context):

$$g(d) = \begin{cases} g_0, & d = 0, \\ 0, & 0 < d < r_1 \quad \text{or} \quad -r_3 < d < 0, \\ \alpha_1 \cdot \frac{1}{\left(1+\frac{d-r_1}{r_2}\right)^k}, & d \geq r_1, \\ \alpha_2 \cdot \frac{1}{\left(1+\frac{-d-r_3}{r_4}\right)^k}, & d \leq -r_3, \end{cases} \tag{10}$$

where $d \in \mathbb{Z}$, $r_1, r_2, r_3, r_4 \in \mathbb{N}^*$, $r_2 > r_1$, $r_4 > r_3$, $k > 0$, and $0 < \alpha_1, \alpha_2 \leq 1$.

Continuous Optimization via Smoothing To prevent gradient instability at the discrete boundary r_1 during backpropagation, we introduce a scaled Sigmoid smoothing kernel $\sigma_s(x)$:

$$\sigma_s(x) = \sigma(2\pi x) = \frac{1}{1 + e^{-2\pi x}} \in (0, 1). \tag{11}$$

Integrating $\sigma_s(x)$ yields our continuous **Single-Shell Model (AGF-S)** for Decoder architectures:

$$g_{\text{smooth}}(d) = \begin{cases} g_0, & d = 0, \\ \sigma_s(d - r_1 + 0.5) \cdot \frac{1}{\left(1+\frac{d-r_1}{r_2}\right)^k}, & d > 0, \end{cases} \tag{12}$$

ensuring smooth differentiability while strictly preserving maximum affinity near the boundary r_1.

5 Experiments and Empirical Results

5.1 Experimental Overview

We evaluate our proposed models on two standard machine translation benchmarks: WMT17 English–German (En–De) and WMT16 English–Romanian (En–Ro). All models are implemented

within the OpenNMT framework (Klein et al., 2017). Relative positional encodings are incorporated strictly within self-attention modules, leaving cross-attention layers unconstrained. Detailed model architectures, optimization dynamics, and parameter bounds are provided in Section 7.

5.2 Positional Curve Analysis in AGF-S

To empirically validate whether our learned parameters align with syntactic expectations, we extract and visualize the learned positional affinity curves $g(d)$ of **AGF-S** across various layers for both En–De and En–Ro benchmarks. Figure 3 illustrates a representative multi-head positional profile from Layer 4.

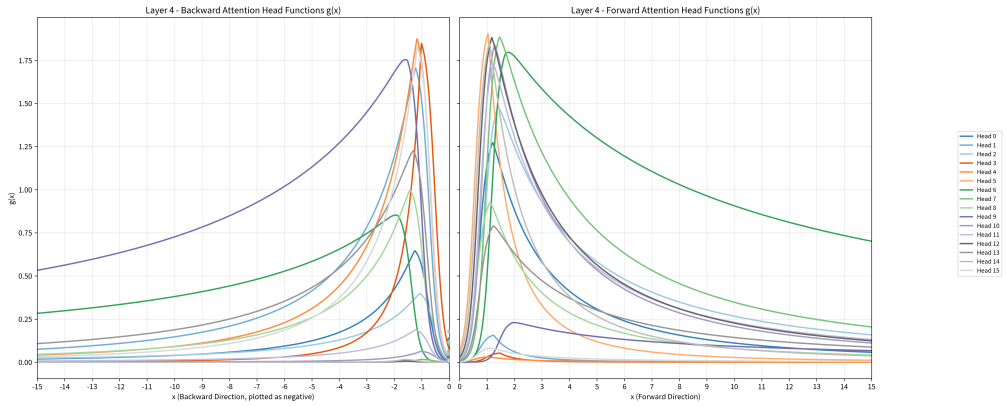

Figure 3: Learned positional affinity curves across distinct attention heads in Layer 4 on WMT17 En–De, showing head specialization in inner shell radius (r_1) and directional scaling.

Comprehensive analysis across all transformer layers reveals several key patterns:

- **Inner Shell Radius (r_1) Localization:** For the vast majority of attention heads, the learned inner shell radius tightly converges near $r_1 \approx 1.0$, with a minority settling near $r_1 \approx 2.0$ or within the continuous interval $(1.0, 2.0)$. This confirms that functional attention pairing in translation predominantly models short-range syntactic dependencies ($d \leq 2$).
- **Directional Asymmetry:** A significant majority of attention heads display distinct peak magnitudes between forward (left context) and backward (right context) directions ($\alpha_1 \neq \alpha_2$). This empirically verifies that individual heads specialize directionally to process distinct syntactic structures.
- **Self-Attention Behavior (g_0 at $d = 0$):** Consistent with our linguistic analysis, g_0 converges near zero ($g_0 \approx 0$) across most heads, with only a minor fraction exhibiting moderate values ($0.3 \leq g_0 \leq 0.5$). This strongly supports our *Self-Attention Zero-Distance Paradox* hypothesis.

5.3 Multi-Shell Extensions (AGF-MS)

Empirical visualizations indicating learned inner shell radii continuously spanning $r_1 \in [1, 2]$ highlight a limitation of the single-shell model: syntactic constraints in natural language cannot be fully captured by a single rigid boundary. To address this, we extend **AGF-S** into two distinct Multi-Shell formulations (**AGF-MS**), grounded in physical mass superimposition and linguistic distance invariance, respectively.

Physically Grounded Multi-Shell (AGF-MS-Phys) Under Newton's Spherical Shell Theorem, the external gravitational force exerted by a uniform spherical shell is independent of its radius and determined solely by its enclosed mass. For n concentric spherical shells with inner radii $i \in \{1, \ldots, n\}$ and normalized mass fractions α_i, the total gravitational field superposition is formulated

Table 1: Test BLEU scores on WMT16 En–Ro and WMT17 En–De tasks. Best results in **bold**.

Model Variant	En–Ro (BLEU)	En–De (BLEU)
Vanilla-Base (Vaswani et al., 2017)	33.0	35.2
ALiBi (Press et al., 2022)	32.7	35.0
ALiBi-L (Learnable)	32.9	35.1
log@KERPLE (Chi et al., 2022)	32.8	35.0
AGF-S (Ours)	33.1	35.2
AGF-MS-Ling (Ours)	**33.2**	**35.3**
AGF-MS-Phys (Ours)	**33.2**	**35.3**

as:

$$g_i(d) = \begin{cases} 0, & d < i, \\ \frac{1}{\left(1+\frac{d}{r_2}\right)^k}, & d \geq i, \end{cases} \quad g(d) = \begin{cases} g_0, & d = 0, \\ \sum_{i=1}^{n} \alpha_i \, g_i(d), & d \geq 1, \end{cases} \tag{13}$$

where $d, i, r_2 \in \mathbb{Z}^+$, $r_2 > i > 0$, $k > 0$, $\alpha_i \geq 0$, and $\sum_{i=1}^{n} \alpha_i = 1$.

Linguistically Grounded Multi-Shell (AGF-MS-Ling) In natural language, intervening structural elements (e.g., auxiliary verbs, prepositions, or subword splits) may expand the minimal syntactic distance from $d = 1$ to $d = 2$ without fundamentally altering the relative spatial decay rate of the core modifier-head relationship. This structural shift implies that the relative decay ratio beyond the shell remains invariant for $d \geq 3$: $\frac{g_1(d)}{g_1(d+1)} = \frac{g_2(d+1)}{g_2(d+2)}$. Incorporating this spatial shift yields the linguistic multi-shell kernel:

$$g_i(d) = \begin{cases} 0, & d < i, \\ \frac{1}{\left(1+\frac{d-i}{r_2}\right)^k}, & d \geq i. \end{cases} \tag{14}$$

Since empirical observations show no active syntactic dependencies at $r_1 \geq 3$, we set $n = 2$ in our benchmark implementations, initializing mixture proportions to $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$.

5.4 Overall Performance

We evaluate machine translation BLEU scores on WMT16 En–Ro and WMT17 En–De test sets in Table 1. Our Multi-Shell model (**AGF-MS**) consistently outperforms the standard Sinusoidal baseline, whereas ALiBi variants and log@KERPLE underperform the baseline by 0.2–0.3 BLEU. Overall, **AGF-MS** surpasses prior relative decay mechanisms by 0.3–0.5 BLEU.

5.5 Model Analysis and Ablations

Single-Shell Dynamics: AGF-S vs. AGF **AGF-S** achieves a steady $+0.2$ BLEU gain over unconstrained **AGF**. Individual component ablations show:

- **Self-Attention Bias (g_0):** Fixing $g_0 \in \{0, 1\}$ yields negligible variance (≤ 0.1 BLEU), matching visual evidence that g_0 naturally converges close to zero.
- **Directional Scaling (α_1, α_2):** Constraining directionality ($\alpha_1 = \alpha_2 = 1$) results in a subtle drop (0.10–0.15 BLEU), suggesting heads prioritize intra-directional decay over cross-directional scale matching.
- **Inner Radius Void (r_1):** Omitting r_1 causes an approximate drop of 0.1 BLEU, consistent with learned r_1 values clustering near 1.0.

Multi-Shell Benefits: AGF-MS vs. AGF-S Transitioning to **AGF-MS** brings an additional $+0.1$ BLEU improvement. Superimposing multiple shells resolves boundary ambiguities when peak affinity weights alternate between $r_1 = 1$ and $r_1 = 2$.

Decay Function Trade-offs In our machine translation experiments, ALiBi-L (exponential decay) performs competitively with power-law variants (< 0.1 BLEU difference). Exponential decay exhibits steeper initial drop-offs compared to heavy-tailed power-law functions (see Figure 4). We speculate that this rapid attenuation sharpens local attention at the potential cost of long-range context—a characteristic that may provide a favorable trade-off for translation tasks that naturally prioritize localized syntactic structures.

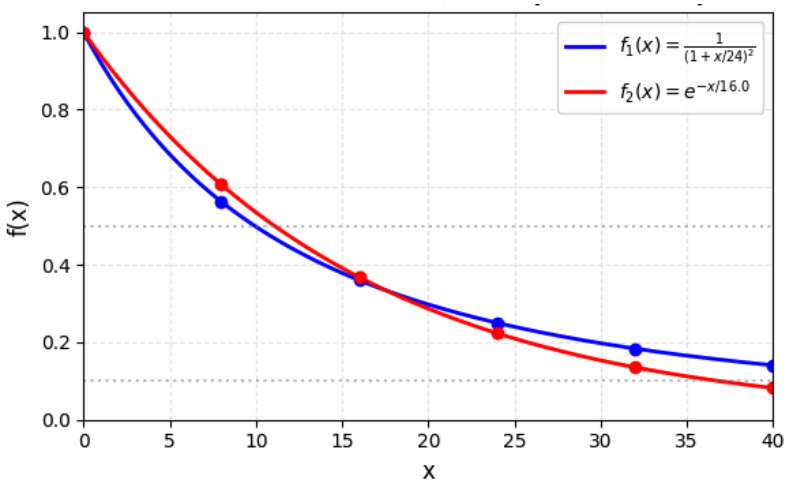

Figure 4: Comparative attenuation trajectories between heavy-tailed power-law decay (AGF/log@KERPLE) and rapid exponential decay (ALiBi).

However, as demonstrated in the KERPLE paper (Chi et al., 2022), when evaluated across broader datasets and larger model capacities, power-law variants (such as log@KERPLE and our proposed AGF) show slight but consistent empirical advantages over exponential variants. We hypothesize that the heavy-tailed property of power-law decay mitigates over-penalization of distant token interactions. As sequence lengths and model scales expand, this persistence of long-range context appears to become increasingly critical for capturing distant semantic dependencies.

6 Limitations

While the AGF-MS framework establishes a physically grounded and mathematically rigorous positional encoding, its empirical validation is currently limited to standard machine translation benchmarks (WMT17 En–De and WMT16 En–Ro) due to computational constraints. Large-scale pretraining on multi-billion parameter LLMs across multi-terabyte corpora remains to be fully executed.

7 Conclusion and Future Extensions

In this paper, we re-examine relative positional representation in Transformer architectures by establishing a theoretical bridge between linguistic syntax and attention mechanics. Through multi-faceted theoretical derivation, we hypothesize that syntactic positional affinity follows a power-law decay profile, uncovering a striking mathematical convergence with Newton's Universal Law of Gravitation and the Spherical Shell Theorem. By integrating core linguistic constraints, we refine this physical analogy from a continuous spherical model into discrete Single-Shell (**AGF-S**) and Multi-Shell (**AGF-MS**) formulations. Empirical evaluations across multi-lingual translation benchmarks confirm that AGF-MS consistently outperforms prior baselines. The theoretical framework established here offers a principled reference point for future architectural design and mechanistic interpretability studies.

AI USE STATEMENT

In this work, generative AI models and language assistance tools were utilized solely for text editing, grammatical corrections, and language refinement. All primary theoretical derivations, model architectures, experimental setups, and quantitative evaluations were developed entirely by the authors, who assume full responsibility for the final content.

ETHICS STATEMENT

This paper focuses on theoretical positional representation in Transformer models. The work does not involve sensitive human data, social discrimination risks, or direct potential for malicious misuse, and strictly complies with ethical guidelines.

REPRODUCIBILITY STATEMENT

To guarantee full reproducibility, all experimental hyperparameter setups, optimization constraints, and evaluation details are summarized below:

- **Datasets & Framework:** All models are implemented in OpenNMT (Klein et al., 2017). For WMT16 En–Ro, we follow the preprocessed split from Lee et al. (2018) (610k train, 2k valid, 2k test). For WMT17 En–De, standard preprocessing is applied. Evaluation relies on `sacreBLEU` (Post, 2018) with `tokenize=none`.
- **Architectures & Training Setup:** En–De uses Transformer-Big ($h = 16, d_{\text{model}} = 1024$); En–Ro uses Transformer-Base ($h = 8, d_{\text{model}} = 512$). The initial learning rate is set to 2.0 with 4000 warmup steps using the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.998$). All models are trained on NVIDIA A100 GPUs until convergence.

REFERENCES

Ta-Chung Chi, Ting-Hao Huang, and Alexander I. Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 8386–8399, 2022.

J. T. Duane. Learning curve approach to reliability monitoring. Technical Report Technical Report Technical Information Series No. R64MD191, General Electric Company, 1964.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, 2017.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. Deterministic non-autoregressive neural sequence generation by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1182, 2018.

Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, pp. 186–191, 2018.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 464–468, 2018.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

George Kingsley Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, 1935.