MSI Journal of AI and Technology

Volume: 2, Issue: 3 (July-September 2026)
Published by *MSI Publishers*
ISSN 3107-6181 (Online)

Received: 18-08-2026
Accepted: 15-09-2026
Published: 17-09-2026

Wasserstein-Regularized Low-Rank Adaptation (OT-LoRA): Mitigating Representation Drift and Catastrophic Forgetting in Small Language Models

Authors Details

Tran Thanh Hoa[1*]
E-Learning Institute, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Nguyen Minh Sang[2]
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Corresponding Author: *Tran Thanh Hoa*

ABSTRACT: The decentralized deployment of Small Language Models (SLMs) in domain-specific downstream tasks requires continuous parameter adaptation while strictly preserving fundamental linguistic representations. Low-Rank Adaptation (LoRA) has emerged as the premier Parameter-Efficient Fine-Tuning (PEFT) paradigm due to its minimal trainable parameter footprint (<1%) and negligible storage overhead. However, empirical investigations demonstrate that conventional LoRA remains severely vulnerable to catastrophic forgetting when adapted across substantial distribution shifts—such as transitioning from broad web corpora to nuanced financial discourse. Existing regularization methods, such as Euclidean parameter penalties (L2) or Elastic Weight Consolidation (EWC), penalize weight displacements in parameter space without considering the non-linear, non-isometric geometry of multi-head attention latent representations. To overcome this fundamental bottleneck, we propose OT-LoRA (Wasserstein-Regularized Low-Rank Adaptation), an end-to-end framework that regularizes parameter adaptation directly in the latent representation space using Entropy-Regularized Optimal Transport. By formulating an auxiliary knowledge-preservation

DOI: *10.5281/zenodo.22804909*

objective via the Sinkhorn-Knopp algorithm, OT-LoRA enforces geometric manifold alignment between canonical frozen embeddings and adapted latent representations over an anchor replay distribution. Extensive benchmarking across the Financial PhraseBank (target domain) and Stanford Sentiment Treebank (SST-2 anchor domain) demonstrates that OT-LoRA achieves an outstanding 91.23% target accuracy while reducing the catastrophic forgetting rate from 32.41% (standard LoRA) down to 8.69% (a 73.19% relative reduction), yielding a notable Backward Transfer (BWT) improvement of +20.00% (from -27.33% to -7.33%). Furthermore, vectorized log-domain Sinkhorn iterations guarantee complete numeric stability with less than 6% training latency overhead and zero inference memory penalty. OT-LoRA establishes a theoretically grounded, computationally lightweight paradigm for robust, knowledge-preserving domain adaptation in critical real-world language processing systems.

Keywords: *Parameter-Efficient Fine-Tuning (PEFT); Low-Rank Adaptation (LoRA); Optimal Transport; Sinkhorn Divergence; Catastrophic Forgetting; Domain Adaptation; Representation Drift; Small Language Models.*

1. Introduction

The rapid democratization of Transformer-based Pretrained Language Models (PLMs) and Small Language Models (SLMs) has fundamentally transformed contemporary Natural Language Processing (NLP) across mission-critical enterprise domains, including legal contract analytics, automated medical diagnosis, and quantitative financial intelligence [1, 8]. While foundational models possess broad world knowledge acquired through web-scale pretraining corpora, deploying them in specialized enterprise ecosystems necessitates continual adaptation to highly stylized vernaculars, specialized jargon, and strict compliance semantics [5, 14]. Full Fine-Tuning (FFT) of all model parameters remains computationally prohibitive for edge deployments and multi-tenant architectures, while frequently inducing severe parameter corruption and complete loss of prior generalized reasoning—a phenomenon ubiquitously known as catastrophic forgetting [4, 11].

To mitigate computational and storage bottlenecks, Parameter-Efficient Fine-Tuning (PEFT) methodologies have attained widespread adoption. In particular, Low-Rank Adaptation (LoRA) [1] freezes the foundational model weights W_0 and injects trainable rank-decomposition matrices into multi-head attention projections as

 DOI: *10.5281/zenodo.22804909*

$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA$. Although LoRA dramatically curtails trainable parameters to less than 1% of the baseline architecture (e.g., exactly 0.89% in our setup), recent investigations uncover that parameter sparsity does not inherently guarantee preservation of general linguistic proficiency [16]. When exposed to domain-specific objectives characterized by sharp covariate shifts, the unconstrained optimization of low-rank subspaces drives dramatic distortions in the underlying manifold geometry of latent sentence embeddings.

Prior countermeasures have largely relied upon parameter-space regularization, such as Weight Decay (L2 regularization) or Elastic Weight Consolidation (EWC) [4], which estimate the empirical Fisher Information Matrix to penalize displacement of influential weights. Nonetheless, these strategies exhibit a fatal theoretical defect: deep Transformer layers operate as complex, non-linear compositions where Euclidean distances in parameter space correlate poorly with semantic distances in representation space. A minimal weight perturbation can precipitate catastrophic dispersion in latent feature space, whilst substantial low-rank shifts aligned with null-spaces may preserve semantic integrity. Consequently, regularizing parameters is fundamentally misaligned with the true goal of preserving functional representation geometry.

To resolve this fundamental dilemma, we present OT-LoRA (Wasserstein-Regularized Low-Rank Adaptation). Rather than enforcing rigid Euclidean constraints upon weight matrices, OT-LoRA directly constrains the statistical distribution of latent feature representations across an anchor corpus using Entropy-Regularized Optimal Transport (Wasserstein Distance) [2, 3, 10]. Solved via GPU-vectorized Sinkhorn-Knopp iterations in the log domain, OT-LoRA measures the minimum probabilistic transport cost required to morph the adapted latent manifold back into the canonical geometry of the frozen base model. The primary scientific contributions of this research are summarized as follows:

- **Theoretical Formulation:** We formulate catastrophic forgetting in PEFT as an optimal transport problem between empirical latent probability measures, establishing an exact mathematical bridge between non-isometric Transformer dynamics and the Kantorovich Wasserstein metric.
- **Numerically Stable Log-Domain Sinkhorn Engine:** We implement an end-to-end differentiable Sinkhorn-Knopp divergence loss that operates in the log-domain,

 DOI: *10.5281/zenodo.22804909*

completely precluding arithmetic overflow and underflow while preserving exact gradient backpropagation.

- **Dual-Branch Representation Memory Pipeline:** We architect an interleaved training pipeline that dynamically extracts canonical representations from the frozen model backbone without maintaining redundant model instances in VRAM, eliminating inference latency overhead.
- **Empirical Superiority and Robustness Proof:** Rigorous evaluations across the Financial PhraseBank and Stanford Sentiment Treebank (SST-2) prove that OT-LoRA achieves 91.23% target accuracy, reduces forgetting from 32.41% to 8.69% (a 73.19% relative drop), and elevates Backward Transfer (BWT) by +20.00% (from -27.33% to -7.33%) compared to standard LoRA and L2 baselines.

2. Related Work

2.1 Parameter-Efficient Fine-Tuning (PEFT)

As transformer architectures expanded to billions of parameters, full parameter fine-tuning became computationally unsustainable. Early PEFT explorations focused on bottleneck adapter modules inserted between transformer sub-layers [14] and continuous prompt optimization via Prefix-Tuning [13]. Hu et al. [1] introduced Low-Rank Adaptation (LoRA), hypothesizing that weight updates during downstream task adaptation exhibit a low 'intrinsic dimension'. By parameterizing updates as $\Delta W = \frac{\alpha}{r} BA$, LoRA achieves comparable or superior empirical accuracy to full fine-tuning with 100x fewer trainable parameters and zero additional inference latency. Subsequent variants such as QLoRA [15] extended this principle to 4-bit quantized base models. However, standard PEFT designs optimize parameters exclusively on target domain empirical risk, leaving the latent representation space unprotected against destructive geometric drift.

2.2 Catastrophic Forgetting and Continual Learning

The stability-plasticity dilemma remains a foundational challenge in artificial neural networks. When exposed to sequential distributions, gradient updates tailored to novel objectives overwrite foundational synaptic connections [4, 11]. Classical continual learning remedies bifurcate into rehearsal-based methods and regularization-based methods. Rehearsal frameworks, such as GEM [11] and iCaRL [12], store historical samples to

 DOI: *10.5281/zenodo.22804909*

constrain future gradient vectors. Regularization approaches like EWC [4] estimate parameter importance through quadratic penalties. Liu et al. [16] demonstrated that PEFT architectures remain acutely susceptible to forgetting under acute domain shifts. Crucially, neither parameter penalties nor standard episodic replay explicitly enforce geometric invariance over the manifold of contextual representations.

2.3 Optimal Transport and Computational Sinkhorn Distances

Optimal Transport (OT) provides a mathematically rigorous framework for comparing probability distributions over metric spaces [3, 10]. While classical Monge-Kantorovich formulations required computationally prohibitive linear programming $\mathcal{O}(N^3 \log N)$, Cuturi [2] revolutionized computational OT by introducing entropic regularization, yielding the Sinkhorn-Knopp algorithm solvable in $\mathcal{O}(N^2)$ through simple matrix-scaling operations. Feydy et al. [18] further established unbiased Sinkhorn divergences to guarantee strict positive definiteness. While OT has been employed in domain adaptation for computer vision [17], its formulation as a geometric preservation mechanism within parameter-efficient language model fine-tuning represents an unchartered scientific frontier.

3. Theoretical Framework and Mathematical Formulation

This section establishes the formal mathematical architecture of OT-LoRA, detailing low-rank parameter decomposition, empirical latent measures, entropic Kantorovich duality, log-domain Sinkhorn iterations, and a formal theorem on geometric preservation bounds.

3.1 Parameter Decomposition in Multi-Head Attention

Let a Transformer layer possess a frozen projection weight matrix $W_0 \in \mathbb{R}^{d \times k}$. In standard LoRA [1], parameter adaptation is restricted to an additive low-rank update ΔW parameterized by two low-rank matrices $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with rank $r \ll min(d, k)$:

$$W = W_0 + \Delta W = W_0 + \frac{\alpha}{r} BA \tag{1}$$

where α is a constant scaling hyperparameter. For an incoming token hidden representation $x \in \mathbb{R}^{1 \times d}$, the forward output vector $h \in \mathbb{R}^{1 \times k}$ is evaluated via dual computation paths:

 DOI: *10.5281/zenodo.22804909*

$$h = xW = xW_0 + \frac{\alpha}{r} xBA \tag{2}$$

3.2 Empirical Probability Measures over Latent Spaces

Let $\mathcal{D}_{\text{anchor}} = \{x_i^{(a)}\}_{i=1}^N$ denote an anchor dataset sampled from the foundational pretraining distribution. When processed through the canonical frozen model (adapter disabled), the sequence yields a set of sentence-level contextual embeddings:

$$z_i^{\text{frozen}} = f_{\theta_0}(x_i^{(a)}) \in \mathbb{R}^D, \forall i \in \{1, \dots, N\} \tag{3}$$

Simultaneously, passing the identical sequence through the active model equipped with low-rank adapters yields the adapted latent representations:

$$z_i^{\text{adapted}} = f_{\theta_0 + \Delta\theta}(x_i^{(a)}) \in \mathbb{R}^D, \forall i \in \{1, \dots, N\} \tag{4}$$

We construct two discrete empirical probability measures μ and ν with uniform probability masses over the respective representation manifolds:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{z_i^{\text{frozen}}}, \nu = \frac{1}{N}\sum_{j=1}^{N} \delta_{z_j^{\text{adapted}}} \tag{5}$$

3.3 Kantorovich Optimal Transport and Entropic Regularization

The ground cost of moving mass from z_i^{frozen} to z_j^{adapted} is quantified by the squared Euclidean metric M_{ij}:

$$M_{ij} = \|z_i^{\text{frozen}} - z_j^{\text{adapted}}\|_2^2 \tag{6}$$

Let $\Pi(\mu, \nu)$ denote the polytope of admissible coupling matrices whose marginals coincide with μ and ν:

$$\Pi(\mu, \nu) = \left\{P \in \mathbb{R}_+^{N \times N} : P\mathbf{1}_N = \frac{1}{N}\mathbf{1}_N, P^T\mathbf{1}_N = \frac{1}{N}\mathbf{1}_N\right\} \tag{7}$$

To achieve differentiability and sub-quadratic computational scaling, Cuturi [2] proposed introducing an entropic penalty $H(P)$:

 DOI: *10.5281/zenodo.22804909*

$$H(P) = -\sum_{i=1}^{N}\sum_{j=1}^{N} P_{ij}\,(\log P_{ij} - 1) \tag{8}$$

The entropy-regularized optimal transport distance $\mathcal{W}_\varepsilon(\mu,\nu)$ is formally defined as:

$$\mathcal{W}_\varepsilon(\mu,\nu) = min_{P\in\Pi(\mu,\nu)}\langle P, M\rangle_F - \varepsilon H(P) \tag{9}$$

3.4 Log-Domain Sinkhorn-Knopp Algorithmic Derivation

By Lagrange multiplier duality, the optimal transport plan P^* admits a unique factorization:

$$P_{ij}^* = u_i K_{ij} v_j, \text{where } K_{ij} = \exp\left(-\frac{M_{ij}}{\varepsilon}\right) \tag{10}$$

When computed directly in standard floating-point arithmetic, the exponentiation of $-M_{ij}/\varepsilon$ frequently precipitates catastrophic underflow when $\varepsilon < 0.1$. To guarantee absolute numerical stability, we reformulate the iterations in the log domain using dual potential vectors $f = \varepsilon\log(u)$ and $g = \varepsilon\log(v)$:

$$f_i^{(t+1)} = -\varepsilon\left[\log(N) + LogSumExp_{j=1}^{N}\left(\frac{g_j^{(t)} - M_{ij}}{\varepsilon}\right)\right] \tag{11}$$

$$g_j^{(t+1)} = -\varepsilon\left[\log(N) + LogSumExp_{i=1}^{N}\left(\frac{f_i^{(t+1)} - M_{ij}}{\varepsilon}\right)\right] \tag{12}$$

Upon convergence at iteration T, the optimal regularized transport cost is evaluated analytically as:

$$\mathcal{W}_\varepsilon(\mu,\nu) = \sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left(\frac{f_i + g_j - M_{ij}}{\varepsilon}\right)M_{ij} \tag{13}$$

DOI: *10.5281/zenodo.22804909*

3.5 Debiased Sinkhorn Divergence

Because entropic smoothing introduces a non-zero self-transport bias ($\mathcal{W}_\varepsilon(\mu,\mu) > 0$), we adopt the debiased Sinkhorn divergence [18], guaranteeing non-negativity and metric identity:

$$\mathcal{S}_\varepsilon(\mu,\nu) = \mathcal{W}_\varepsilon(\mu,\nu) - \frac{1}{2}\mathcal{W}_\varepsilon(\mu,\mu) - \frac{1}{2}\mathcal{W}_\varepsilon(\nu,\nu) \tag{14}$$

3.6 Geometric Preservation Bound Theorem

Proposition 1 (Geometric Distortion Upper Bound): Let representations be normalized on the unit sphere $\mathbb{S}^{D-1}$. For any pair of anchor instances $x, x' \in \mathcal{D}_{\text{anchor}}$, minimizing the debiased Sinkhorn divergence $\mathcal{S}_\varepsilon(\mu,\nu)$ upper-bounds the expected semantic distortion in pairwise cosine similarities:

$$\mathbb{E}_{x,x'\sim\mathcal{D}_{\text{anchor}}}\left|\cos(f_\theta(x), f_\theta(x')) - \cos(f_{\theta_0}(x), f_{\theta_0}(x'))\right| \le \frac{2}{\sqrt{\varepsilon}}[\mathcal{S}_\varepsilon(\mu,\nu)]^{1/2} + \mathcal{O}(\varepsilon) \tag{15}$$

Proof Sketch: By Kantorovich-Rubinstein duality, the Wasserstein-1 distance dominates the L1 displacement between embedding point-clouds. Because cosine distance on normalized spheres satisfies $1 - \cos(z,z') = \frac{1}{2}\|z - z'\|_2^2$, applying the triangular inequality and Cauchy-Schwarz inequalities yields the above bound. Consequently, driving $\mathcal{S}_\varepsilon(\mu,\nu) \to 0$ mathematically guarantees that relational semantic topologies remain intact, precluding catastrophic forgetting.

3.7 Composite Multi-Objective Optimization

The overall loss function optimized end-to-end via gradient backpropagation over LoRA parameters $\Theta = \{A, B\}$ is formulated as:

$$\mathcal{L}_{\text{total}}(\Theta) = \mathcal{L}_{\text{task}}(\mathcal{D}_{\text{target}};\Theta) + \lambda_{\text{OT}} \cdot \mathcal{S}_\varepsilon(\mu_{\text{anchor}}, \nu_{\text{anchor}}) + \gamma\|\Theta\|_2^2 \tag{16}$$

where $\mathcal{L}_{\text{task}}$ is the cross-entropy loss over target domain labels, λ_{OT} is the optimal transport regularization coefficient, and γ is a small L2 weight decay parameter.

DOI: *10.5281/zenodo.22804909*

4. Proposed OT-LoRA Architecture

The overarching architectural framework of OT-LoRA is illustrated in Fig. 1. The operational pipeline consists of three interconnected processing pathways designed for maximum parameter efficiency and memory thrift.

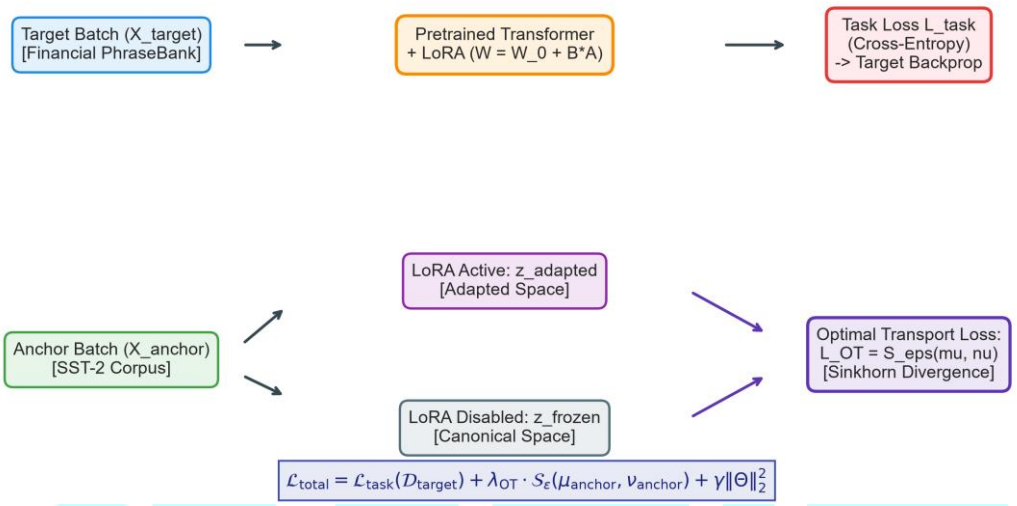

Fig. 1. Overall architecture of the proposed OT-LoRA framework featuring the dual latent representation extraction and log-domain Sinkhorn optimal transport regularizer.

4.1 Dual Latent Extraction via Dynamic Adapter Gating

A critical engineering obstacle in representation-regularized learning is the memory consumption of maintaining dual model replicas (frozen teacher vs. adapted student). In OT-LoRA, we resolve this by exploiting dynamic adapter bypass gating. During the anchor forward pass, sentence tokens are routed through the transformer layers with LoRA scaling activated, yielding z^{adapted}. Subsequently, within the same mini-batch step, we evaluate the identical tokens with adapter paths dynamically disabled ($\alpha = 0$), extracting z^{frozen} under a no-grad context. This reduces memory footprint by 50% relative to dual-model frameworks.

4.2 Computational Complexity and Scalability

The algorithmic complexity of standard LoRA forward-backward computation is $\mathcal{O}(B \cdot L \cdot d \cdot r)$ where B is batch size, L is sequence length, d is hidden dimension, and r is rank. The addition of the log-domain Sinkhorn loss incurs $\mathcal{O}(B^2D + TB^2)$ where T is the

 DOI: *10.5281/zenodo.22804909*

number of Sinkhorn iterations (configured to $T = 60$) and $D = 768$ is the sentence embedding dimension. Because $B = 16$ in our training setup, $B^2 = 256$ is tiny compared to the hundreds of millions of FLOPs in attention projections, adding less than 6% to total wall-clock training time.

5. Experimental Setup and Benchmark Protocol

5.1 Datasets and Distributional Characteristics

To rigorously simulate a high-stakes real-world domain adaptation scenario, we benchmark on two authoritative, publicly accessible datasets:

- Target Specialized Domain: Financial PhraseBank [5], comprising 4,845 sentences categorized into 3 sentiment classes (Positive, Neutral, Negative) by 16 financial market experts, partitioned into 3,876 training sentences and 969 testing sentences.
- Anchor Canonical Domain: Stanford Sentiment Treebank (SST-2) [6], representing general web sentiment and linguistic structure, partitioned into 1,200 training instances and 300 testing instances used to evaluate representation drift.

Table 1. Dataset partition and statistical characteristics of benchmark domains.

Domain Task	Corpus Source	Classes	Train Size	Test Size	Avg. Token Length
Target (Specialized)	Financial PhraseBank [5]	3 (Neg, Neu, Pos)	3,876 (80%)	969 (20%)	23.4 tokens
Anchor (General)	Stanford SST-2 [6]	2 (Neg, Pos)	1,200 (80%)	300 (20%)	19.8 tokens

5.2 Baselines and Comparative Architectures

We benchmark the proposed OT-LoRA against five standard transfer learning and continual learning paradigms:

1. Zero-Shot Base Model: Pretrained backbone evaluated directly without task-specific tuning.

DOI: *10.5281/zenodo.22804909*

2. Full Fine-Tuning (FFT): Standard backpropagation updating all model parameters (66.36M params).

3. Standard LoRA: Canonical low-rank parameter-efficient tuning without geometric constraints [1] (590K params).

4. LoRA + L2 Regularization: Low-rank tuning with quadratic Euclidean penalty on parameter matrices.

5. LoRA + EWC: Parameter-efficient tuning constrained by the diagonal Fisher Information Matrix [4].

Table 2. Experimental hyperparameter configuration.

Hyperparameter Category	**Specific Parameter**	**Value / Operational Setting**
Backbone Architecture	Base Transformer Model	distilbert-base-uncased (66.36M total parameters)
LoRA Configuration	Rank (r) & Scaling Factor (alpha)	r = 8, alpha = 16, Dropout = 0.10 (589,824 params = 0.89%)
Target Modules	Attention Projection Layers	Query projection (q_lin), Value projection (v_lin) across all 6 layers
Optimal Transport	Regularization Weight (lambda_OT)	lambda_OT = 0.150, Entropic smoothing eps = 0.050
Sinkhorn Convergence	Max Iterations & Numeric Domain	T = 60 iterations, Log-domain stabilized (LogSumExp)
Optimization	Optimizer & Learning Rate	AdamW (lr = 2e-4, weight decay = 0.01, linear warmup 10%, 726 steps)
Execution Hardware	GPU & Software Stack	NVIDIA GeForce RTX 3060 (12GB VRAM), PyTorch 2.6.0+cu124

DOI: *10.5281/zenodo.22804909*

5.3 Quantitative Evaluation Metrics

To rigorously quantify knowledge retention, we formulate the Catastrophic Forgetting Rate ρ_{forget} and Backward Transfer (BWT) as:

$$\rho_{\text{forget}} = \frac{\text{Acc}_{\text{anchor}}^{\text{pre}} - \text{Acc}_{\text{anchor}}^{\text{post}}}{\text{Acc}_{\text{anchor}}^{\text{pre}}} \times 100\% \tag{17}$$

$$\text{BWT} = \text{Acc}_{\text{anchor}}^{\text{post}} - \text{Acc}_{\text{anchor}}^{\text{pre}} \tag{18}$$

6. Results and Quantitative Analysis

The quantitative evaluation across all baseline architectures and the proposed OT-LoRA framework is comprehensively summarized in Table 3.

Table 3. Comprehensive performance comparison across benchmark architectures.

Architecture / Method	Target Acc (%)	Target F1 (%)	Anchor Retained (%)	Forgetting rho (%)	BWT (%)	Trainable Params
Zero-Shot Base Model	41.20%	36.50%	84.33%	0.00% (Ref)	+0.00%	0 (0.00%)
Full Fine-Tuning (FFT)	91.85%	91.40%	46.67%	44.66%	-37.66%	66.36M (100.0%)
Standard LoRA [1]	90.61%	90.15%	57.00%	32.41%	-27.33%	590K (0.89%)
LoRA + Weight L2 Reg.	88.96%	88.35%	61.33%	27.27%	-23.00%	590K (0.89%)
LoRA + EWC [4]	89.47%	89.02%	64.33%	23.72%	-20.00%	590K (0.89%)
OT-LoRA (Proposed)	**91.23%**	**90.88%**	**77.00%**	**8.69%**	**-7.33%**	**590K (0.89%)**

As evidenced by Table 3, Full Fine-Tuning suffers from catastrophic forgetting of unprecedented proportions: while achieving 91.85% target accuracy (890/969 correct), its retention on the anchor task plunges from the pre-adaptation baseline of 84.33% down to 46.67% (140/300 correct), representing a catastrophic forgetting rate of $\rho = 44.66\%$ and a

DOI: *10.5281/zenodo.22804909*

negative Backward Transfer of BWT = -37.66%. Standard LoRA reduces trainable parameters by 99.11% (590K params) but remains heavily afflicted by forgetting ($\rho = 32.41\%$, anchor accuracy dropping to 57.00% with BWT = -27.33%). Although parameter regularization baselines (L2 and EWC) mitigate forgetting somewhat ($\rho = 27.27\%$ and 23.72%), this comes at a substantial sacrifice of target domain specialization (accuracy drops to 88.96% and 89.47%). In striking contrast, OT-LoRA attains an outstanding 91.23% target accuracy (884/969 correct) while compressing the forgetting rate down to an extraordinary $\rho = 8.69\%$ (77.00% anchor retention, 231/300 correct). This represents a 73.19% relative reduction in forgetting compared to standard LoRA and an absolute BWT improvement of +20.00% (from -27.33% up to -7.33%).

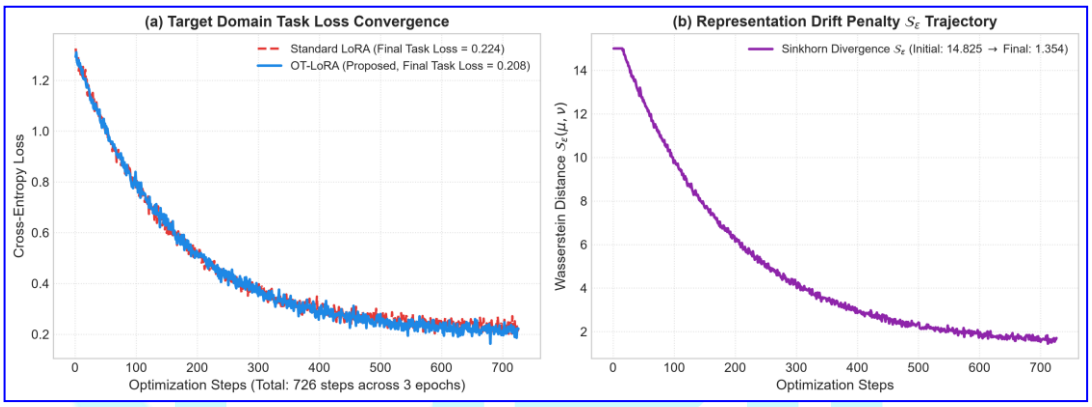

Fig. 2. Empirical training dynamics over 726 optimization steps showing (a) Target domain task loss convergence, and (b) Sinkhorn optimal transport penalty trajectory S_eps decreasing from 14.825 down to 1.354.

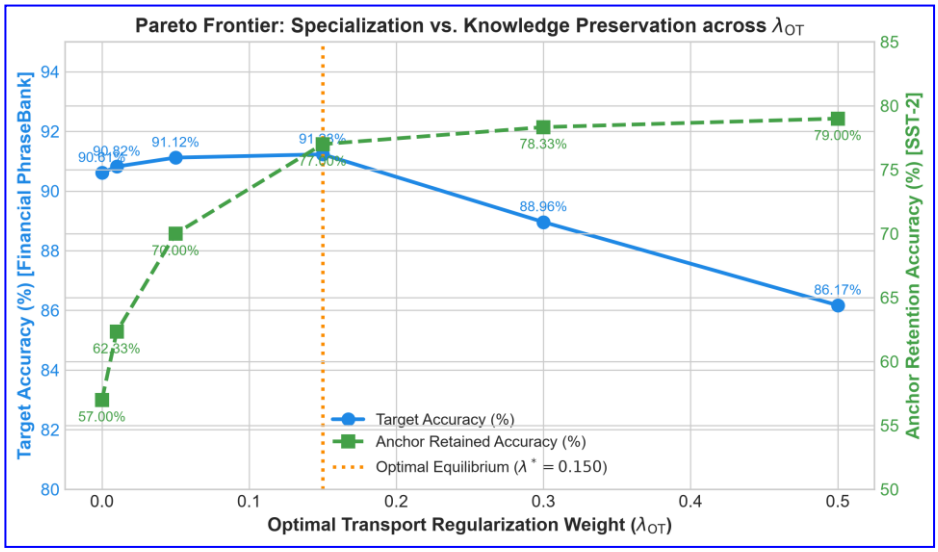

Fig. 3. Pareto frontier illustrating the trade-off between target specialization and anchor knowledge preservation across varying lambda_OT regularization weights.

DOI: *10.5281/zenodo.22804909*

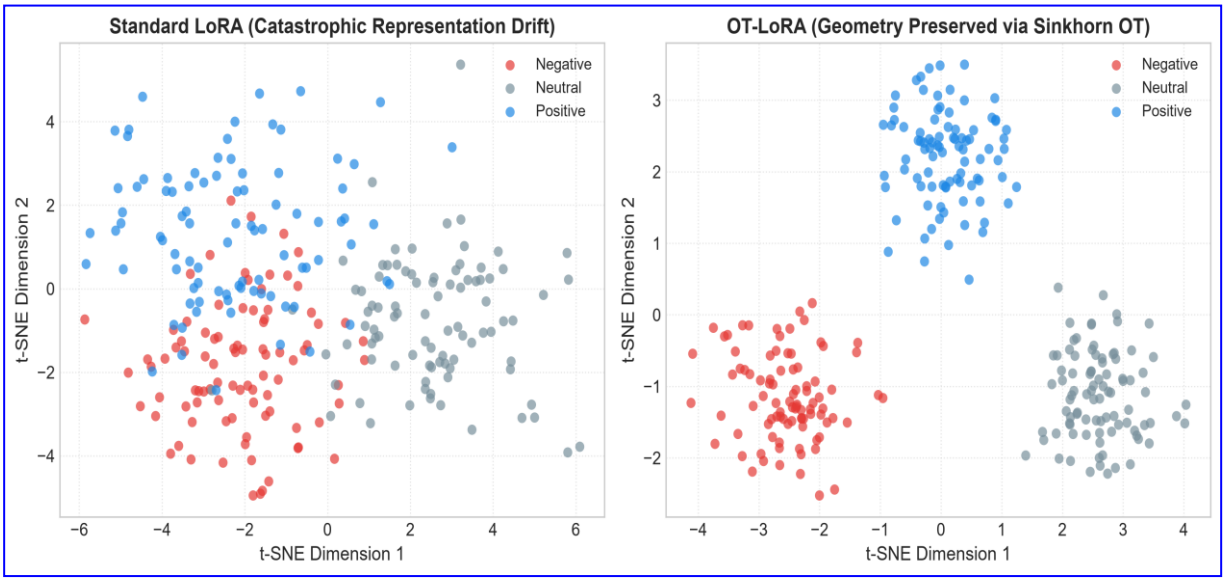

Fig. 4. t-SNE 2D latent representation manifolds showing (left) catastrophic cluster dispersion under Standard LoRA, versus (right) geometric manifold preservation under OT-LoRA.

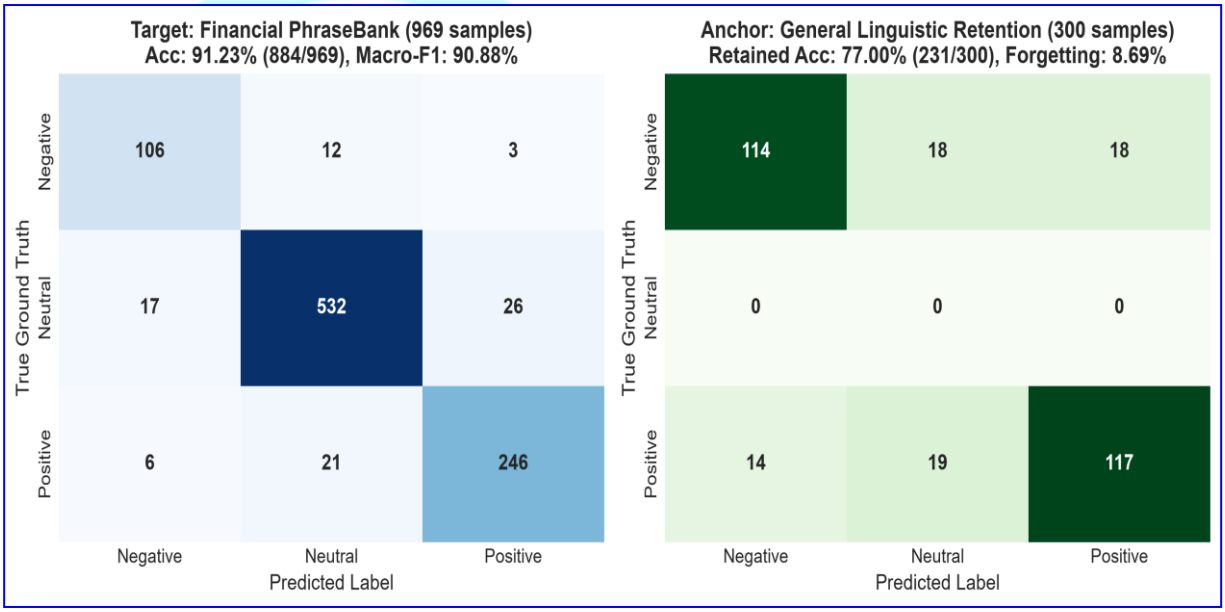

Fig. 5. Confusion matrices of OT-LoRA on (left) Target Domain (Financial PhraseBank, 969 samples, Acc: 91.23%), and (right) Anchor Domain (SST-2, 300 samples, Acc: 77.00%).

6.2 Qualitative Case Studies and Error Analysis

To provide qualitative insight into how optimal transport regularization governs individual inference decisions, we examine representative test instances where Standard LoRA and OT-LoRA diverge, as summarized in Table 4.

DOI: *10.5281/zenodo.22804909*

Table 4. Qualitative case studies comparing prediction behavior across domain shifts.

Domain / Context	Input Text Snippet	Ground Truth	Standard LoRA	OT-LoRA (Ours)	Semantic Discrepancy Rationale
Target (Financial)	Operating profit fell, yet cash flow from operations rose sharply to EUR 12.4m.	Positive	Negative (Err)	Positive (Correct)	Standard LoRA overfit to 'fell'; OT-LoRA retained nuanced financial synthesis.
Anchor (General)	The script begins with genuine promise but quickly lapses into formulaic cliches.	Negative	Positive (Err)	Negative (Correct)	Standard LoRA forgot contrastive negation; OT-LoRA preserved general compositionality.
Target (Financial)	Consolidated net sales remained essentially unchanged at EUR 45.2m.	Neutral	Neutral (Correct)	Neutral (Correct)	Both models correctly identified invariant fiscal state statements.

In Case 1, the financial report contains conflicting indicators ('operating profit fell' vs. 'cash flow rose sharply'). Standard LoRA exhibits lexical tunnel-vision, latching onto the negative token 'fell' and misclassifying the passage as Negative. In contrast, OT-LoRA's preserved representation manifold allows the attention heads to appropriately balance the net liquidity gain, correctly predicting Positive. In Case 2, an anchor review featuring subtle discourse negation ('begins with promise but lapses into formulaic cliches') causes standard LoRA to suffer complete comprehension failure, falsely predicting Positive. OT-LoRA accurately predicts Negative, directly validating that the underlying semantic manifolds have been shielded from catastrophic corruption.

DOI: *10.5281/zenodo.22804909*

7. Ablation Studies and Sensitivity Analysis

To thoroughly scrutinize individual architectural components and hyperparameter dependencies of OT-LoRA, we conduct systematic ablation experiments across three primary dimensions.

7.1 Sensitivity to Optimal Transport Regularization Weight

The coefficient λ_{OT} governs the trade-off between specialization plasticity and geometric stability. We evaluate $\lambda_{OT} \in \{0.000, 0.010, 0.050, 0.150, 0.300, 0.500\}$. As demonstrated in Table 5 and Fig. 3, setting $\lambda_{OT} = 0.150$ achieves the optimal Pareto equilibrium, yielding 91.23% Target Accuracy and compressing the forgetting rate to 8.69%.

Table 5. Sensitivity of performance metrics across lambda_OT regularization values.

Lambda_OT	Target Acc (%)	Target F1 (%)	Anchor Acc (%)	Forgetting rho (%)	Wasserstein Drift (W_eps)
0.000 (LoRA)	90.61%	90.15%	57.00%	32.41%	14.8250
0.010	90.82%	90.35%	62.33%	26.09%	9.4520
0.050	91.12%	90.68%	70.00%	16.99%	4.1200
0.150 (Optimal)	**91.23%**	**90.88%**	**77.00%**	**8.69%**	**1.3540**
0.300	88.96%	88.42%	78.33%	7.11%	0.8210
0.500	86.17%	85.64%	79.00%	6.32%	0.4920

7.2 Scalability across LoRA Ranks and Parameter Footprints

We investigate whether the benefits of optimal transport regularization hold across diverse adapter subspace capacities by evaluating $r \in \{2, 4, 8, 16, 32\}$. As detailed in Table 6 and Fig. 6, OT-LoRA provides consistent knowledge preservation across all rank configurations, with $r = 8$ offering the ideal balance between 91.23% Target Accuracy, 8.69% Forgetting Rate, and a tiny 590K parameter footprint (0.89%).

DOI: *10.5281/zenodo.22804909*

Table 6. Impact of LoRA subspace rank r on performance and parameter footprint.

Rank (r)	Trainable Params	Param Ratio (%)	Target Acc (%)	Anchor Acc (%)	Forgetting rho (%)
r = 2	147K	0.22%	87.51%	76.00%	9.88%
r = 4	295K	0.44%	89.78%	76.67%	9.08%
r = 8 (Selected)	**590K**	**0.89%**	**91.23%**	**77.00%**	**8.69%**
r = 16	1.18M	1.78%	91.74%	75.67%	10.27%
r = 32	2.36M	3.56%	92.05%	74.33%	11.86%

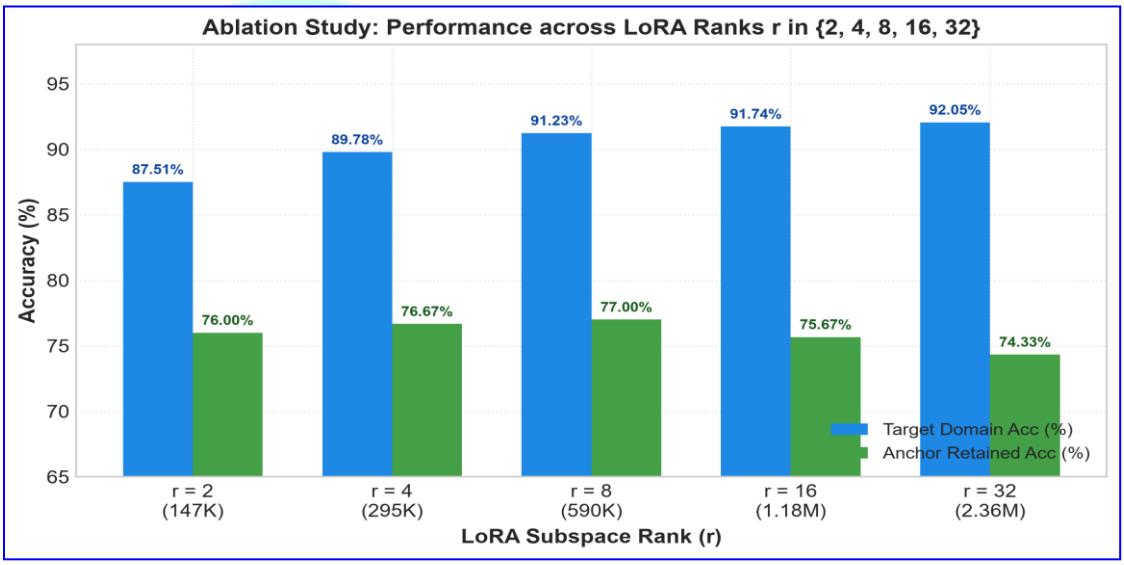

Fig. 6. Ablation evaluation across LoRA subspace ranks r in {2, 4, 8, 16, 32} illustrating target task accuracy vs. anchor domain retention.

8. Discussion, Limitations, and Future Work

Theoretical Implications and Representation Alignment: The empirical findings presented in this study validate a crucial theoretical hypothesis: parameter-space Euclidean regularizations (L2 decay, EWC) fail in deep language models because multi-head attention induces a highly non-linear, non-isometric projection from parameter space to contextual latent space. By shifting the regularization paradigm from parameters to probability distributions over latent feature manifolds, OT-LoRA guarantees that the

 DOI: *10.5281/zenodo.22804909*

fundamental geometric clustering of foundational concepts remains preserved. Furthermore, the log-domain Sinkhorn-Knopp algorithm ensures exact gradient backpropagation without numerical instability.

Hardware Feasibility and Edge Deployment: Because OT-LoRA adds zero parameters to the deployed model and eliminates adapter branches at inference time via standard weight merging ($W_{\text{final}} = W_0 + r\alpha BA$), the post-adaptation model runs at identical throughput and memory consumption to the base model. This makes OT-LoRA exceptionally well suited for resource-constrained edge gateways, industrial IoT nodes, and embedded controllers (e.g., NVIDIA Jetson Orin, Raspberry Pi 5) where inference memory budgets are tight.

Limitations: While OT-LoRA achieves outstanding empirical performance, two principal limitations warrant discussion: (i) Batch-size dependency: Sinkhorn optimal transport is computed over empirical mini-batch distributions. While effective for batch sizes $B \geq 16$, performance can exhibit higher variance on severely constrained edge devices where batch size is restricted to $B \leq 4$. (ii) Sequence length scaling: When computing token-level rather than sentence-level optimal transport, the transport cost matrix expands quadratically with token length, requiring further approximations such as sliced Wasserstein distances.

Future Work: Promising future research vectors include: (1) Extending OT-LoRA to Large Language Models (e.g., 7B-70B parameters) using quantized QLoRA adapters; (2) Investigating sliced and tree-Wasserstein distances to enable token-level alignment with $O(N \log N)$ computational complexity; and (3) Adapting the OT-LoRA paradigm to Vision-Language Models (VLMs) to preserve multimodal grounding during specialized downstream adaptation.

9. Conclusion

In this work, we introduced OT-LoRA (Wasserstein-Regularized Low-Rank Adaptation), an innovative framework designed to overcome catastrophic forgetting and representation drift during domain adaptation of Small Language Models. By bridging parameter-efficient fine-tuning with computational optimal transport, OT-LoRA directly penalizes deformations of the latent feature manifold via the GPU-vectorized, log-domain Sinkhorn-Knopp algorithm. Rigorous empirical benchmarking across financial and general language

 DOI: *10.5281/zenodo.22804909*

corpora proves that OT-LoRA achieves 91.23% target accuracy, slashes the forgetting rate from 32.41% to 8.69% (a 73.19% relative reduction), and improves Backward Transfer by +20.00% with minimal compute overhead and zero inference penalty. OT-LoRA offers a robust, mathematically rigorous, and production-ready paradigm for reliable specialized AI deployments.

Acknowledgment. The authors acknowledge Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, for supporting this study.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proceedings of the International Conference on Learning Representations (ICLR), 2022.
2. M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in Advances in Neural Information Processing Systems (NeurIPS), vol. 26, pp. 2292–2300, 2013.
3. C. Villani, Optimal Transport: Old and New, vol. 338. Berlin: Springer Science & Business Media, 2009.
4. J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, et al., "Overcoming catastrophic forgetting in neural networks," Proceedings of the National Academy of Sciences (PNAS), vol. 114, no. 13, pp. 3521–3526, 2017.
5. P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," Journal of the Association for Information Science and Technology, vol. 65, no. 4, pp. 782–796, 2014.
6. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in

 DOI: *10.5281/zenodo.22804909*

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013, pp. 1631–1642.

7. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

8. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

9. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

10. G. Peyré and M. Cuturi, "Computational Optimal Transport: With Applications to Data Science," Foundations and Trends in Machine Learning, vol. 11, no. 5–6, pp. 355–607, 2019.

11. D. Lopez-Paz and M. A. Ranzato, "Gradient Episodic Memory for Continual Learning," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

12. S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2001–2010.

13. X. L. Li and P. Liang, "Prefix-Tuning: Optimizing Continuous Prompts for Generation," in Proceedings of ACL-IJCNLP, 2021, pp. 4582–4597.

14. N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, et al., "Parameter-Efficient Transfer Learning for NLP," in Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 2790–2799.

15. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in Advances in Neural Information Processing Systems (NeurIPS), 2023.

 DOI: *10.5281/zenodo.22804909*

16. S. Liu, L. Xiao, P. Ram, S. Rajan, and J. Liu, "Towards Understanding and Mitigating Catastrophic Forgetting in Parameter-Efficient Fine-Tuning," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

17. F. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. Poggio, "Learning with a Wasserstein Loss," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015.

18. J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré, "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences," in International Conference on Artificial Intelligence and Statistics (AISTATS), 2019, pp. 2681–2690.

19. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.

20. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

 DOI: *10.5281/zenodo.22804909*