

Grokking and epoch-wise double descent in quantum neural networks

Daniel Pranjic^{1,2}, Marco Roth¹, and Christian Tutschku¹

¹Fraunhofer IAO, Nobelstraße 12, 70569 Stuttgart, Germany

²Universität Stuttgart, Allmandring 35, 70569 Stuttgart, Germany

Abstract

Grokking, the delayed transition from memorization to generalization, is a fundamental phenomenon in gradient-based learning, yet its dynamics within variational quantum machine learning (QML) remain largely unexamined. In this work, we report the empirical observation of both the grokking transition and epoch-wise double descent in a two-qubit quantum neural network (QNN) under a complete parameterization of the $SU(4)$ manifold. We demonstrate that overparameterization via increased circuit depth improves the probability of successful generalization. Notably, these architectures frequently exhibit an epoch-wise double descent in test error, degrading at a critical epoch before recovering into a generalizing state. Crucially, we identify a *generalization decay* in late-stage training, where the test error increases significantly despite a stagnant training loss. Bridging this behavior with algorithmic stability theory, our analysis reveals that this decay correlates with an unconstrained increase of the weight-norm, drifting away from sparse, phase-aligned harmonic solutions toward overfitted solutions in the Hilbert space. We analyze the underlying temporal dynamics of this transition, demonstrating how the onset of generalization is linked to optimization hyperparameters such as learning rate and weight decay. Finally, to mitigate late-stage decay, we introduce a weak explicit weight-norm regularization into the loss function. We demonstrate that this structural anchor stabilizes the post-grokking phase and permanently preserves generalization gains, providing a robust framework for training overparameterized quantum circuits.

1 Introduction

Understanding the phase transitions of overparameterized models remains a central challenge in machine learning. A hallmark of these dynamics is **grokking**: a sharp shift from memorization to generalization occurring long after training performance saturates [1]. While classical studies analyze this via weight-norms and singular learning theory, recent work suggests grokking may manifest as an entanglement transition [2], hinting at a deeper quantum-mechanical dimension that necessitates a systematic translation to true quantum architectures. However, while the quantum machine learning (QML) community heavily focuses on trainability and Barren Plateaus [3], the dynamics of Quantum Neural Networks (QNNs) *after* loss convergence remain largely unexplored. In overparameterized regimes [4], optimization continues along flat, degenerate valleys of the loss landscape structurally linked to double descent phenomena [5], yet QNNs often lack the structural robustness to permanently maintain these late-stage generalization gains.

In this work, we demonstrate that QNNs fundamentally exhibit both grokking and epoch-wise double descent, driven by the non-Euclidean geometry of the compact, periodic quantum gate manifold. Our findings **challenge the prevailing dogma surrounding Barren Plateaus**, which traditionally mandates narrow parameter initializations ($\sigma_\theta \rightarrow 0$) to preserve gradients. We demonstrate that in the overparameterized deep-circuit regime, broad initializations paired with strong weight decay do not trap the model; instead, they actively facilitate a highly structured representation search, compressing the timeline to generalization. Thus, overparameterization is a vital resource providing the parameter degrees of freedom for implicit regularization to systematically filter out non-generalizing solutions during training.

Primary contributions

- 1) Observation of quantum grokking (Subsection 3.1):** We provide the first empirical evidence of the grokking transition in $SU(4)$ QNNs using a minimal 2-qubit architecture to track landscape traversal, showing that the training plateau is a dynamic regime where the model drifts across regions of varying algorithmic stability, Lipschitz bounds and entanglement.
- 2) Statistical universality (Subsection 3.2):** We demonstrate via multi-seed analysis that quantum grokking is an intrinsic feature of overparameterized architectures rather than a rare anomaly, showing that increased circuit depth systematically improves the success rate of generalization.
- 3) Temporal dynamics (Subsection 3.3):** We quantify the “time until generalization”, revealing how learning rates and implicit regularization control the onset of the generalization transition.

2 Setup and theory

The goal is to find a controlled minimal setting that shows the grokking transition and epoch-wise double descent while connecting it to algorithmic stability theory.

2.1 Dataset and margin-based testing

Following the approach in [6], we construct a non-linear concentric two-class task to evaluate the quality of the learned QNN decision boundary. The training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ consists of points sampled randomly from two concentric circles with labels $y^{(i)} \in \{-1, 1\}$ (see Figure 1a). To specifically probe the model’s structural robustness, the test set $\mathcal{D}_{\text{test}}$ consists of points distributed evenly at a small margin $\varepsilon > 0$ along the optimal decision boundary, which is pre-calculated using a classical hard-margin SVM with a quadratic kernel [7].

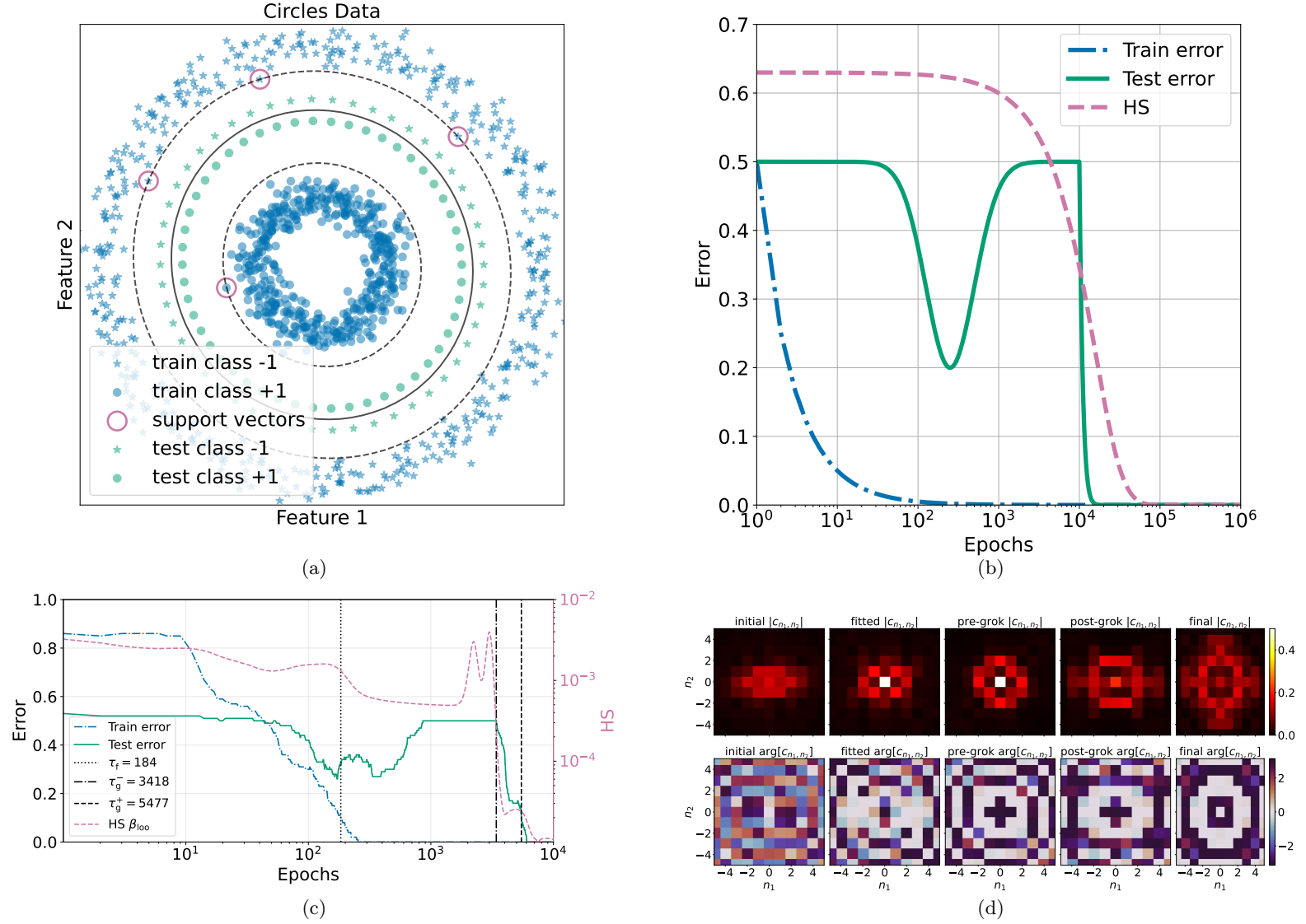


Figure 1: (a) Blueprint of concentric training data. (b) Idealized quantum grokking sketch. (c) Empirical optimization history. (d) Output function Fourier magnitudes $|c_{n1, n2}|$ alignment tracker.

2.2 Overparameterized QNNs

To isolate training dynamics on the $SU(4)$ manifold without non-linear data wrapping, we embed 2D samples into a 15D vector $\phi(\mathbf{x}) = (x_1, x_2, 1, \dots, 1) \in \mathbb{R}^{15}$. Each layer l of our n_l -layer QNN applies an $SU(4)$ unitary parameterized by the 15 Lie algebra generators $\{G_k\}_{k=1}^{15}$ of the form $U^{(l)}(\mathbf{x}, \boldsymbol{\theta}) = \exp(i \sum_{k=1}^{15} \theta_k^{(l)} \phi_k(\mathbf{x}) G_k)$, with weights sampled from a Gaussian $\theta_k^{(l)} \sim \mathcal{N}(0, \sigma_\theta)$. The model output is the expectation value $f(\mathbf{x}; \boldsymbol{\theta}) = \langle \psi(\mathbf{x}, \boldsymbol{\theta}) | M | \psi(\mathbf{x}, \boldsymbol{\theta}) \rangle$ for measurement $M = Z \otimes Z$ and state $|\psi(\mathbf{x}, \boldsymbol{\theta})\rangle = \prod_{l=1}^{n_l} U^{(l)}(\mathbf{x}, \boldsymbol{\theta}^{(l)}) |00\rangle$. We minimize the regularized MSE loss using AdamW [8] with implicit regularization strength $\lambda_W > 0$ and $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{train}}} (y - f(\mathbf{x}; \boldsymbol{\theta}))^2 + \lambda \sum_{k=1}^{15} \sum_{l=1}^{n_l} (\theta_k^{(l)})^2$, where $\lambda \geq 0$ is the global weight-norm regularization strength [9].

2.3 Algorithmic stability theory

To formally characterize the generalization dynamics during the grokking transition, we utilize distribution-agnostic algorithmic stability theory [6, 10]. The expected generalization gap is upper-bounded by the *uniform* hypothesis stability (HS) β , such that the expected risk $\mathcal{R}$ and empirical risk $\mathcal{R}_{\text{emp}}$ satisfy $\mathbb{E}[\mathcal{R}] \leq \mathbb{E}[\mathcal{R}_{\text{emp}}] + \beta$. An algorithm $A_{\mathcal{H}}$ governed by hyperparameters $\mathcal{H}$, which outputs a predictor f from a hypothesis family, exhibits uniform HS β if the absolute change in loss incurred by omitting any single training sample z_i from dataset $\mathcal{D} \in \mathcal{Z}^N$ is bounded across all evaluation points $z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ by $\sup_{\mathcal{D}, z, i} |\ell(A_{\mathcal{H}}(\mathcal{D}), z) - \ell(A_{\mathcal{H}}(\mathcal{D} \setminus \{z_i\}), z)| \leq \beta$. Because evaluating a pointwise supremum over non-linear QNN landscapes is computationally intractable, we track the expected *leave-one-out* HS over the data distribution ($\beta_{100} \leq \beta$), defined as

$$\beta_{100} = \mathbb{E}_{\mathcal{D}, z} [|\ell(A_{\mathcal{H}}(\mathcal{D}), z) - \ell(A_{\mathcal{H}}(\mathcal{D} \setminus \{z_i\}), z)|]. \quad (1)$$

In our empirical evaluations, we explicitly compute and track β_{100} , though more efficient proxies exist that skip leave-one-out retraining, such as analyzing the spectral properties of the representation Gramian [6, 11].

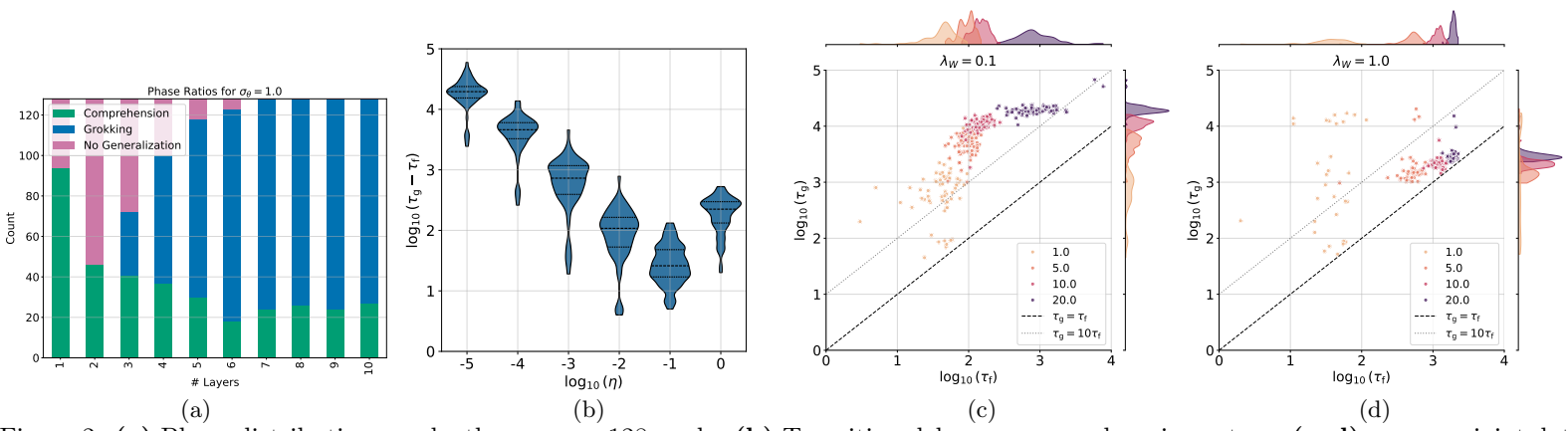


Figure 2: (a) Phase distribution vs. depth n_l across 128 seeds. (b) Transition delay $\tau_g - \tau_f$ vs. learning rate η . (c–d) τ_g vs. τ_f jointplot with varying initialization spread σ_θ for weak ($\lambda_W = 0.1$) and strong ($\lambda_W = 1.0$) weight decay configurations.

3 Highlighted results

We analyze our findings through three lenses: [Subsection 3.1](#) details a concrete grokking transition and epoch-wise double descent within a single canonical training run, [Subsection 3.2](#) presents a multi-seed statistical analysis demonstrating that grokking is an intrinsic feature of QNNs whose success rate systematically scales with circuit depth and [Subsection 3.3](#) characterizes the temporal scaling properties governing the onset of these generalization transitions.

3.1 Empirical observation of the grokking transition and epoch-wise double descent

As shown in [Figure 1c](#), the QNN trajectory progresses from initial fitting ($\tau_f = 184$, training error < 0.1 , test error ~ 0.3) through a static memorization plateau ($\tau_f \rightarrow \tau_g^- = 3418$) to a sudden grokking transition ($\tau_g^+ = 5477$, test error < 0.1). During the seemingly stationary plateau, hidden internal parameter reorganization is captured by the dynamic evolution of the hypothesis stability β_{100} , which drops sharply precisely at the transition. To expose the underlying mechanics, [Figure 1d](#) tracks the QNN output’s Fourier coefficients $c_{n_1, n_2}(\theta_\tau)$. Diffuse and chaotic at $\tau = 0$, the magnitudes $|c_{n_1, n_2}|$ cluster centrally at fitting (τ_f) using low-frequency modes for memorization. Crucially, during the plateau ($\tau_f \rightarrow \tau_g^-$), coefficients evacuate this central density to form an outer ring, revealing that weight decay and gradient dynamics actively filter out non-generalizing modes before the macroscopic transition. At grokking (τ_g^+), this sparse topology stabilizes into a symmetric ring with phase arguments resolved into highly ordered, concentric bands that persist through $\tau = 10^5$. This geometric alignment mirrors the ideal radially symmetric step target function $f(\mathbf{x}) = 2 \cdot \mathbb{I}(r \leq R) - 1$ (R being the radius of the ideal boundary), whose discrete Fourier transform yields $c_{0,0} \approx 2\pi R^2/L^2 - 1$ at the origin and $c_{n_1, n_2} \approx 2RJ_1(2\pi Rn/L)/Ln$, $n = \sqrt{n_1^2 + n_2^2}$ otherwise (J_1 being the first-order Bessel function). This landscape is dominated by an Airy-like diffraction pattern with concentric rings decaying asymptotically as $\mathcal{O}(\|\mathbf{n}\|^{-3/2})$, explaining the ring emergence in [Figure 1d](#). Because this target is purely real and centered, its theoretical coefficients must be real, forcing a binary phase profile (0 or $\pm\pi$). The post-grokking phase bands directly visualize the QNN successfully collapsing chaotic parameter overlays into the low-frequency manifold defined by the alternating roots of the classical Bessel function.

3.2 Occurrence of the grokking phase and generalization success rate

To systematically investigate grokking occurrence and stability, we conduct a multi-seed analysis initializing the QNN across 128 independent random seeds ($\sigma_\theta = 1.0$) while varying circuit depth $n_l \in \{1, \dots, 10\}$. We categorize training trajectories into three distinct phases using a threshold of 10% error: *Comprehension* (train and test errors drop below threshold with a relative time gap $\tau_g/\tau_f \leq 10$); *Grokking* (both drop below threshold with a delayed transition $\tau_g/\tau_f > 10$); both Memorization (only train error drops below threshold) and Confusion (neither error drops below threshold) with individual generalization failures consolidated under the label *No Generalization*. Our statistical results, illustrated in [Figure 2a](#), reveal a striking architectural dependence on QNN depth. For shallow instances ($n_l < 3$), grokking is completely absent within 10^5 epochs, with trajectories split between clean comprehension and generalization failure. However, increasing depth beyond $n_l = 2$ correlates with both the grokking occurrence rate and global generalization success. For deep circuits ($n_l \geq 7$), initialization failure modes are completely suppressed. This proves that overparameterization via layer depth fundamentally smoothens the optimization landscape, mitigating unfavorable initial seed trajectories [\[4, 12\]](#).

3.3 Temporal dynamics of the grokking transition

We define the fit time τ_f and generalization time τ_g as the epochs where training and test errors first drop below 0.1, respectively [\[13\]](#). Our results indicate that explicit regularization strength λ does not alter the temporal distribution of the relative gap τ_g/τ_f ; instead, the foundational driver of quantum grokking remains the implicit regularization from the AdamW optimizer’s decoupled weight decay. As shown in [Figure 2b](#), increasing the learning rate from $\log_{10}(\eta) = -5$ to -1 monotonically accelerates generalization, pulling the median delay ($\tau_g - \tau_f$) down by three orders of magnitude. At $\log_{10}(\eta) = 0$, this trend inverts as extreme gradient noise introduces severe optimization instability. Furthermore, [Figure 2c](#) and [Figure 2d](#) track the joint impact of initialization scale σ_θ and weight decay λ_W . Under weak weight decay ($\lambda_W = 0.1$), models with broad initializations ($\sigma_\theta \geq 10.0$) cluster in the upper-right regime, fitting training data late ($\log_{10}(\tau_f) > 2$) and delaying grokking toward the training horizon. Increasing weight decay to $\lambda_W = 1.0$ ([Figure 2d](#)) compresses these trajectories and accelerates τ_g , shifting high- σ_θ clusters downward. However, marginal distributions reveal lower total counts for $\sigma_\theta \in \{10, 20\}$ under $\lambda_W = 1.0$, proving that aggressive weight decay chokes out global success rates for extreme seeds. This trade-off refines the traditional QML dogma mandating narrow initializations ($\sigma_\theta \rightarrow 0$) to avoid Barren Plateaus: deep, broadly initialized QNNs can successfully generalize via implicit weight decay, provided the regularization is balanced to prevent over-constraining the parameter manifold.

References

- [1] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.
- [2] Domenico Pomarico, Alfonso Monaco, Giuseppe Magnifico, Antonio Lacalamita, Ester Pantaleo, Loredana Bellantuono, Sabina Tangaro, Tommaso Maggipinto, Marianna La Rocca, Ernesto Picardi, et al. Grokking as an entanglement transition in tensor network machine learning. *arXiv preprint arXiv:2503.10483*, 2025.
- [3] Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018.
- [4] Martin Larocca, Nathan Ju, Diego García-Martín, Patrick J Coles, and Marco Cerezo. Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3(6):542–551, 2023.
- [5] Marie Kempkes, Aroosa Ijaz, Elies Gil-Fuster, Carlos Bravo-Prieto, Jakob Spiegelberg, Evert van Nieuwenburg, and Vedran Dunjko. Double descent in quantum kernel methods. *PRX Quantum*, 7(1):010312, 2026.
- [6] Luca Oneto, Sandro Ridella, Andrea Coraddu, and Davide Anguita. Reconciling grokking with statistical learning theory. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2025.
- [7] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- [8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [9] Julian Berberich, Daniel Fink, Daniel Pranjić, Christian Tutschku, and Christian Holm. Training robust and generalizable quantum models. *Physical Review Research*, 6(4):043326, 2024.
- [10] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.
- [11] Luca Oneto, Sandro Ridella, and Davide Anguita. Do we really need a new theory to understand over-parameterization? *Neuro-computing*, 543:126227, 2023.
- [12] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.
- [13] Zhiyuan Liu, Yankai Lin, and Maosong Sun. *Representation learning for natural language processing*. Springer Nature, 2023.