Contents

Chapter 4

Inductive Knowledge Graph Learning with Active Forgetting

A version of this work was previously presented at a peer-reviewed conference. Please refer to [Chen et al., 2022] for full citation.

Knowledge graphs form the backbone of modern knowledge engines, enabling AI systems to organize, retrieve, and reason over structured information. Among the tools that enrich and sustain these knowledge graphs, Factorization Models (FMs), such as DistMult, have emerged as a cornerstone in Knowledge Graph Completion (KGC), a task focused on predicting missing relationships between entities. In transductive scenarios, Factorization Models (FMs) often surpass Graph Neural Networks (GNNs), emerging as indispensable pillars of knowledge graphs, completing them and elevating their utility as a foundational source of knowledge for downstream tasks.

However, FMs struggle in inductive scenarios, where they can not generalize to unseen nodes or incorporate node features effectively. To transfer FM's transductive performance to inductive scenarios, we observe that FMs' structure formation rely highly on the embeddings. These embeddings, when optimized through gradient descent, can be reinterpreted as a sequence of message-passing rounds across the knowledge graph. In other words, embeddings essentially act as a historical cache of node states, tracing structural traversals over the knowledge graph.

This perspective reveals a fundamental limitation about FMs: when trained to convergence, FMs tend to capture excessive global graph structures through infinite rounds

of implicit message-passing, often far exceeding the graph's natural radius ($L \to \infty$). While extensive structuring yields strong transductive performance, it also results in overly constrained representations that hinder generalization from training graphs to new, unseen graphs. To destructure rigid representations, we propose a simple yet powerful mechanism: *active forgetting*. By periodically clearing and reloading new node embeddings, this operator truncates the infinite rounds of message-passing, resetting the model's memory of past computations over the nodes. This reset forces the model to focus on the local neighbourhood information, which enables inductive reasoning for previously unseen or forgotten nodes. Mathematically, this approach synthesizes the strengths of FMs and GNNs into a unified framework, which we call ReFactor GNNs.

Evaluations across standard KGC benchmarks demonstrate that ReFactor GNNs maintain the transductive performance of FMs while achieving state-of-the-art inductive performance with significantly fewer parameters. ReFactor GNNs bridge the gap between FMs and GNNs, providing a unified architecture for robust knowledge graph representation learning, supporting AI agents' dynamic knowledge needs in the wild.

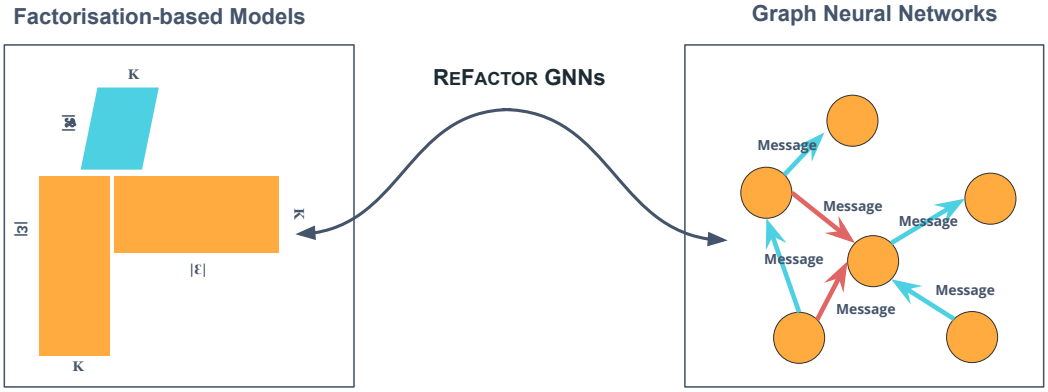

Figure 4.1: ReFactor GNN bridges factorization-based models and graph neural networks by reformulating gradient descents over entity embeddings as message-passing rounds.

4.1 Factorization Meets Message-Passing

In recent years, machine learning on graphs has attracted significant attention due to the abundance of graph-structured data and developments in graph learning algorithms. Graph Neural Networks (GNNs) have demonstrated state-of-the-art performance for many graph-related problems, such as node classification [Kipf and Welling, 2016] and graph classification [Gilmer et al., 2017]. Their main advantage is that they can easily be applied in an inductive setting: generalising to new nodes and graphs without re-training. However, despite many attempts at applying GNNs for multi-relational link prediction tasks such as Knowledge Graph Completion [Nickel et al., 2016c], there are still few positive results compared to more traditional factorisation-based models (FMs) [Yang et al., 2015b, Trouillon et al., 2016]. As it stands, GNNs, after resolving reproducibility concerns, either deliver significantly lower performance [Nathani et al., 2019, Sun et al., 2020a] or yield negligible performance gains at the cost of highly sophisticated architecture designs [Xu et al., 2020b]. A notable exception is NBFNet [Zhu et al., 2021], but even here improvements come at the price of high computational inference costs compared to FMs. Furthermore, it is unclear how NBFNet could incorporate node features, which, as we will see in this work, leads to remarkably lower performance in an inductive setting. On the other hand FMs, despite being a simpler architecture, have been found to be very accurate for knowledge graph completion when coupled with appropriate training strategies [Ruffinelli et al., 2020] and training objectives [Lacroix et al., 2018, Chen et al., 2021]. However, they also come with shortcomings in that they, unlike GNNs, can not be applied in an inductive setting.

Given the respective strengths and weaknesses of FMs and GNNs, we wonder *whether we can bridge these two seemingly different model categories* so that we can develop knowledge graph completion models that generalize to unseen graphs. While exploring this question, we make the following contributions:

- By reformulating gradient descent on node embeddings using message-passing primitives, we show a practical connection between FMs and GNNs, in that: FMs can be treated as a special instance of GNNs, but with infinite neighbourhood, layer-wise training and a global normaliser.[1]

[1]The traditional view is that *the transductive nature of FMs stem from their need to retrain on new*

- Based on this connection, we propose a new family of architectures, referred to as ReFactor GNNs, which interpolates between FMs and GNNs. In essence, ReFactor GNNs inductivise FMs by using a *finite* number of message-passing layers, and incorporating node features.
- Through an empirical investigation across 15 well-established inductive and transductive benchmarks, we find that ReFactor GNNs achieve state-of-the-art inductive performance and comparable transductive performance to FMs, despite using an order of magnitude fewer parameters than GNNs.

4.2 Literature Review: Multi-relational Graph Learning, FMs, and GNNs

Multi-Relational Graph Representation Learning Multi-relational graph representation learning concerns graphs with various edge types. Another relevant line of work would be representation learning over heterogeneous graphs, where node types are also considered. Previous work on multi-relational graph representation learning focused either on FMs [Nickel et al., 2011b, Trouillon et al., 2016, Yang et al., 2015b, Lacroix et al., 2018, Nickel et al., 2016c, Dettmers et al., 2018, Nguyen et al., 2018, Chen et al., 2021] or GNN-based models [Schlichtkrull et al., 2018, Xu et al., 2020a, Zhang et al., 2020, Li et al., 2021b]. Similar to a recent finding in a benchmark study over heterogeneous GNNs [Lv et al., 2021], where the best choices of GNNs for heterogeneous graphs seem to regress to simple homogeneous GNN baselines, the progress of multi-relational graph representation learning also mingles with FMs, the classic multi-relational link predictors. Recently, FMs were found to be significantly more accurate than GNNs for KGC tasks, when coupled with specific training strategies [Ruffinelli et al., 2020, Jain et al., 2020b, Lacroix et al., 2018]. While more advanced GNNs [Zhu et al., 2021] for KBC are showing promise at the cost of extra algorithmic complexity, little effort has been devoted to establishing links between plain GNNs and FMs, which are strong multi-relational link predictors despite their simplicity. Our work aims to *align* GNNs with FMs so that we can combine the strengths from both families of models.

nodes, a view which we further underpin by also observing that *FMs are not inductive due to the need for infinite layers of on-the-fly message-passing.*

Relationships between FMs and GNNs A very recent work [Srinivasan and Ribeiro, 2020] builds a theoretical link between structural GNNs and node (positional) embeddings. However, on one end of the link, the second model category encompasses not merely factorisation-based models but also many practical graph neural networks, between which the connection is unknown. Our work instead offers a more practical link between positional node embeddings produced by FMs and positional node embeddings produced by GNNs, while at the same time focusing on KGC. Beyond FMs in KGC, using graph signal processing theory, Shen et al. [2021] show that matrix factorisation (MF) based recommender models correspond to ideal low-pass graph convolutional filters. They also find infinite neighbourhood coverage in MF although using a different approach and focusing on a different domain in contrast to our work.

Message-passing Message-passing is itself a broad terminology, it is generally discussed under two different contexts. Firstly, as a computational technique, message passing allows recursively decomposing a global function into simple local, parallelisable computations [MacKay, 2003], thus being widely used for solving inference problems in a graphical model. Specifically, we note that message passing-based inference techniques were proposed for matrix completion-based recommendation [Kim et al., 2010] and Bayesian Boolean data decomposition [Ravanbakhsh et al., 2016] in the pre-deep-learning era. Secondly, as a paradigm of parameterising learnable functions over *graph-structured data*, message-passing has recently been used to provide a unified reformulation [Gilmer et al., 2017] for various GNN architectures, including Graph Attention Networks [Veličković et al., 2018], Gated Graph Neural Networks [Li et al., 2016], and Graph Convolutional Networks [Kipf and Welling, 2016]. In this work, we show that FMs can also be cast as a special type of message-passing GNNs by considering the gradient descent updates [Bottou, 2012] over node embeddings as message-passing operations between nodes. To the best of our knowledge, our work is the first to provide such connections between FMs and message-passing GNNs. We show that FMs can be seen as instances of GNNs, with a characteristic feature about the nodes being considered during the message-passing process: our ReFactor GNNs can be seen as using an *Augmented Message-Passing* process on a dynamically re-wired graph [Veličković, 2022].

4.3 Formalizing FMs and GNNs for KGC

Knowledge Graph Completion (KGC) [Nickel et al., 2016b], also known as knowledge base completion (KBC), is a canonical task of multi-relational link prediction. The goal is to predict missing edges given existing edges. Formally, a knowledge graph contains a set of entities (nodes), $\mathcal{E} = \{1, \dots, |\mathcal{E}|\}$, a set of relations (edge types) $\mathcal{R} = \{1, \dots, |\mathcal{R}|\}$, and a set of typed edges between the entities $\mathcal{T} = \{(v_i, r_i, w_i)\}_{i=1}^{|\mathcal{T}|}$, where each triplet (v_i, r_i, w_i) indicates a relationship of type $r_i \in \mathcal{R}$ between the *subject* $v_i \in \mathcal{E}$ and the *object* $w_i \in \mathcal{E}$. Given a node v, we denote its *outgoing* 1-hop neighbourhood as the set of relation-object pairs $\mathcal{N}_+^1[\mathrm{v}] = \{(r, o) \mid (\mathrm{v}, r, o) \in \mathcal{T}\}$, its *incoming* 1-hop neighbourhood as the set of subject-relation pairs $\mathcal{N}_-^1[\mathrm{v}] = \{(r, s) \mid (s, r, \mathrm{v}) \in \mathcal{T}\}$, and its total neighbourhood as the union of the two $\mathcal{N}^1[v] = \mathcal{N}_+^1[v] \cup \mathcal{N}_-^1[v]$. We denote the neighbourhood of v under a specific relation r as $\mathcal{N}_\pm^1[r, v]$. Entities may come with features $X \in \mathbb{R}^{|\mathcal{E}| \times K}$ for describing them, such as textual encodings of their names and/or descriptions. Given a (training) knowledge graph, KGC is evaluated by answering $(v, r, ?)$-style queries i.e. predicting the object given the subject and relation in the triplet. And queries like $(?, r, v')$ are answered using inverse queries $(v', r^{-1}, ?)$ in this work, following [Lacroix et al., 2018].

Following the 1vsAll setting used in Chapter 2 and Ruffinelli et al. [2020], multi-relational link prediction models can be trained via maximum likelihood, by fitting a parameterized conditional categorical distribution $P_\theta(w \mid v, r)$ over the candidate objects of a relation, given the subject v and the relation type r:

$$P_\theta(w|\,\mathrm{v}, \mathrm{r}) = \frac{\exp \Gamma_\theta(\mathrm{v}, \mathrm{r}, w)}{\sum_{u \in \mathcal{E}} \exp \Gamma_\theta(\mathrm{v}, \mathrm{r}, u)} = \mathrm{Softmax}(\Gamma_\theta(\mathrm{v}, \mathrm{r}, \cdot))[w]. \tag{4.1}$$

Here $\Gamma_\theta : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$ is a *scoring function*, which, given a triplet (v, r, w), returns the likelihood that the corresponding edge appears in the knowledge graph.

We illustrate our derivations using DistMult [Yang et al., 2015b] as the score function Γ and defer extensions to general score functions, e.g. ComplEx [Trouillon et al., 2016], to the appendix. For DistMult, the score function Γ_θ is defined as the trilinear dot product of the vector representations corresponding to the subject, relation, and object of the

triplet:

$$\Gamma_\theta(v, r, w) = \langle f_\phi(v), f_\phi(w), g_\psi(r) \rangle = \sum_{i=1}^{K} f_\phi(v)_i f_\phi(w)_i g_\psi(r)_i, \tag{4.2}$$

where $f_\phi : \mathcal{E} \to \mathbb{R}^K$ and $g_\psi : \mathcal{R} \to \mathbb{R}^K$ are learnable maps parameterised by ϕ and ψ that encode entities and relation types into K-dimensional vector representations, and $\theta = (\phi, \psi)$. We will refer to f and g as the entity and relation *encoders*, respectively. If we define the data distribution as $P_D(x) = \frac{1}{|\mathcal{T}|} \sum_{(v,r,w)\in\mathcal{T}} \delta_{(v,r,w)}(x)$, where $\delta_{(v,r,w)}(x)$ is a Dirac delta function at (v, r, w), then the objective is to learn the model parameters θ by minimising the expected negative log-likelihood $\mathcal{L}(\theta)$ of the ground-truth entities for the queries $(v, r, ?)$ obtained from $\mathcal{T}$:

$$\begin{aligned} \arg\min_\theta \mathcal{L}(\theta) \quad \text{where} \quad \mathcal{L}(\theta) &= -\mathbb{E}_{x\sim P_D}[\log(P_\theta(w|v, r)] \\ &= -\frac{1}{|\mathcal{T}|} \sum_{(v,r,w)\in\mathcal{T}} \log P_\theta(w|v, r). \end{aligned} \tag{4.3}$$

During inference, we use P_θ for determining the plausibility of links not present in the training graph.

4.3.1 Factorisation-based Models for KGC

In factorisation-based models, which we assume to be DistMult, the entity encoder f_ϕ and the relation encoder g_ψ are simply parameterised as look-up tables, associating each entity and relation with a continuous distributed representation:

$$f_\phi(v) = \phi[v],\ \phi \in \mathbb{R}^{|\mathcal{E}|\times K} \quad \text{and} \quad g_\psi(r) = \psi[r],\ \psi \in \mathbb{R}^{|\mathcal{R}|\times K}. \tag{4.4}$$

The corresponding score function is then given by

$$\Gamma_\theta(v, r, w) = \langle \phi[v], \phi[w], g(r) \rangle = \sum_{i=1}^{K} \phi[v]_i \phi[w]_i \psi[r]_i. \tag{4.5}$$

4.3.2 GNN-based Models for KGC

GNNs were originally proposed for node or graph classification tasks [Gori et al., 2005, Scarselli et al., 2009]. To adapt them to KGC, previous work has explored two different paradigms: *node-wise entity representations* [Schlichtkrull et al., 2018] and *pair-wise entity representations* [Teru et al., 2020, Zhu et al., 2021]. Though the latter paradigm has shown promising results, it requires computing representations for all pairs of nodes, which can be computationally expensive for large-scale graphs with millions of entities. Additionally, node-wise representations allow for using a single evaluation of $f_\phi(v)$ for multiple queries involving v, resulting in faster batch evaluation.

Models based on the first paradigm differ from pure FMs only in the entity encoder and lend themselves well for a fair comparison with pure FMs. We will therefore focus on this class and leave the investigation of pair-wise representations to future work. Let $q_\phi : \mathcal{G} \times \mathcal{X} \to \bigcup_{S \in \mathbb{N}^+} \mathbb{R}^{S \times K}$ be a GNN encoder, where $\mathcal{G} = \{G \mid G \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}$ is the set of all possible multi-relational graphs defined over $\mathcal{E}$ and $\mathcal{R}$, and $\mathcal{X}$ is the input feature space, respectively. Then we can set $f_\phi(v) = q_\phi(\mathcal{T}, X)[v]$. Following the standard message-passing framework [Gilmer et al., 2017, Battaglia et al., 2018, Hamilton] used by the GNNs, we view $q_\phi = q^L \circ ... \circ q^1$ as the recursive composition of $L \in \mathbb{N}^+$ layers that compute intermediate representations $\{h^l\}$ for $l \in \{1, \ldots, L\}$ with $h^0 = X$ for all entities in the KG. Each layer q^l producing representation h_l is made up of the following three functions:

1. A *message function* $q^l_{\text{M}} : \mathbb{R}^K \times \mathcal{R} \times \mathbb{R}^K \to \mathbb{R}^K$ that computes the message along each edge. Given an edge $(v, r, w) \in \mathcal{T}$, the message function q^l_{M} not only makes use of the node states $h^{l-1}[v]$ and $h^{l-1}[w]$ (as in standard GNNs) but also uses the relation r; denote the message as

$$m^l[v, r, w] = q^l_{\text{M}}\left(h^{l-1}[v], r, h^{l-1}[w]\right);$$

2. An *aggregation function* $q^l_{\text{A}} : \bigcup_{S \in \mathbb{N}} \mathbb{R}^{S \times K} \to \mathbb{R}^K$ that aggregates all messages from the 1-hop neighbourhood of a node; denote the aggregated message as

$$z^l[v] = q^l_{\text{A}}\left(\{m^l[v, r, w] \mid (r, w) \in \mathcal{N}^1[v]\}\right);$$

3. An *update function* $q^l_{\text{U}} : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K$ that produces the new node states h^l by combining previous node states h^{l-1} and the aggregated messages z^l:

$$h^l[v] = q^l_{\text{U}}(h^{l-1}[v], z^l[v]).$$

Different parametrisations of q^l_{M}, q^l_{A}, and q^l_{U} lead to different GNNs. For example, R-GCNs [Schlichtkrull et al., 2018] define the q^l_{M} function using per-relation linear transformations $m^l[v, r, w] = \frac{1}{\mathcal{N}^1[r,v]} W^l_r h^{l-1}[w]$, where W^l_r denotes the weight matrix associated with relation r and $\mathcal{N}^1[r, v]$ represents the degree of v under relation r; q^l_{A} is implemented by a summation and q^l_{U} is a non-linear transformation $h^l[v] = \sigma(z^l[v] + W^l_0 h^{l-1}[v])$, where σ is the sigmoid function. For each layer, the learnable parameters are $\{W^l_r\}_{r\in\mathcal{R}}$ and W^l_0, all of which are matrices in $\mathbb{R}^{K\times K}$. Sometimes applying GNNs over an entire graph might not be feasible due to the size of the graph. Hence, in practice, $f_\phi(v)$ can be approximated with sampled sub-graphs [Hamilton et al., 2017, Zou et al., 2019, Zeng et al., 2020], such as L-hop neighbourhood around node v denoted as $\mathcal{N}^L[v]$:

$$f_\phi(v) = q_\phi(\mathcal{T}_{\mathcal{N}^L[v]}, X_{\mathcal{N}^L[v]})[v]. \tag{4.6}$$

4.4 Implicit Message-Passing in FMs

The sharp difference in analytical forms might give rise to the misconception that GNNs incorporate message-passing over the neighbourhood of each node (up to L-hops), while FMs do not. In this work, we show that by explicitly considering the training dynamics of FMs, we can uncover and analyse the hidden message-passing mechanism within FMs. In turn, this will lead us to the formulation of a novel class of GNNs well suited for multi-relational link prediction tasks (Section 4.5). Specifically, we propose to interpret the FMs' optimisation process of their objective as the entity encoder. After randomly initialising the parameters ϕ of the look-up table, FMs are typically trained to minimise the loss $\mathcal{L}$ (Equation 4.3). If we consider, for simplicity, a gradient descent training dynamic, then the entity encoder operating on a given node v, $f_{\phi^t}(v)$, can be rewritten

as the outcome of a series of gradient descent steps:

$$\begin{aligned} f_{\phi^t}(v) &= \phi^t[v] \\ &= \mathrm{GD}^t(\phi^{t-1}, \mathcal{T})[v] \\ &= \mathrm{GD}^t \circ \mathrm{GD}^{t-1}(\phi^{t-2}, \mathcal{T})[v] \\ &= \underbrace{\mathrm{GD}^t \circ \cdots \circ \mathrm{GD}^1}_{t\text{gradient steps}}(\phi^0, \mathcal{T})[v] \end{aligned} \tag{4.7}$$

where ϕ^t is the embedding vector at the t-th step, $t \in \mathbb{N}^+$ is the total number of training iterations, and ϕ^0 is a random initialisation of the look-up table. GD is the gradient descent operator, which we can expand by substituting in the objective $\mathcal{L}$ (Equation 4.3):

$$\mathrm{GD}(\phi, \mathcal{T}) = \phi - \beta \nabla_\phi \mathcal{L} \tag{4.8}$$

$$= \phi + \alpha \sum_{(v,r,w)\in\mathcal{T}} \frac{\partial \log P(w|v,r)}{\partial \phi}, \tag{4.9}$$

where $\alpha = \beta \left|\mathcal{T}\right|^{-1}$ with a learning rate $\beta > 0$. We now dissect Equation 4.8 in two different but equivalent ways. In the first, which we dub the *edge view*, we separately consider each addend of the gradient $\nabla_\phi \mathcal{L}$. In the second, we aggregate the contributions from all the triplets to the update of a particular node. With this latter decomposition, which we call the *node view*, we can explicate the message-passing mechanism at the core of the FMs. While the edge view suits a vectorised implementation better, the node view further exposes the information flow among nodes, allowing us to draw an analogy to message-passing GNNs.

4.4.1 The Edge View

Each addend of Equation 4.8 corresponds to a single edge $(v, r, w) \in \mathcal{T}$ and contributes to the update of the representation of all nodes. The update on the representation of the

subject v contributed by this edge can be written as:

$$\begin{aligned}
\text{GD}(\phi, \{(v,r,w)\})[v] &= \phi[v] + \alpha \frac{\partial \log P(w|v,r)}{\partial \phi[v]} \\
&= \phi[v] + \alpha \frac{\partial \log \frac{\exp \Gamma(v,r,w)}{\sum_{u \in \mathcal{E}} \exp \Gamma(v,r,u)}}{\partial \phi[v]} \\
&= \phi[v] + \alpha \left(\frac{\partial \Gamma(v,r,w)}{\partial \phi[v]} - \sum_{u \in \mathcal{E}} P(u|v,r) \frac{\partial \Gamma(v,r,u)}{\partial \phi[v]} \right) \\
&= \phi[v] + \alpha \left(\underbrace{g(r) \odot \phi[w]}_{w \to v} \underbrace{- \sum_{u \in \mathcal{E}} P_\theta(u|v,r) g(r) \odot \phi[u]}_{u \to v} \right).
\end{aligned}$$

Step two follows by substituting the softmax expression for the conditional probability (Equation 4.1) and take gradients of the log softmax, where the critical part is the treatment of the gradient of the log partition function:

$$\begin{aligned}
\frac{\partial \log \sum_u \exp(\Gamma(\cdot,u))}{\partial \phi[v]} &= \frac{1}{\sum_u \exp(\Gamma(\cdot,u))} [\sum_u \exp(\Gamma(\cdot,u)) \frac{\partial \Gamma}{\partial \phi[v]}] \\
&= \sum_u \frac{\exp(\Gamma(\cdot,u))}{\sum_u \exp(\Gamma(\cdot,u))} \frac{\partial \Gamma}{\partial \phi[v]} = \sum_u P(u|\cdot) \frac{\partial \Gamma}{\partial \phi[v]}.
\end{aligned}$$

Step three results from taking the gradient of the score function Γ (Equation 4.5):

$$\frac{\partial \Gamma(v,r,w)}{\partial \phi[v]} = \frac{\langle \phi[v], \phi[w], g(r) \rangle}{\partial \phi[v]} = g(r) \odot \phi[w].$$

We discuss the meaning underlying this decomposition. The $w \to v$ term represents information flow from w (a positive neighbour of v) to v, thereby increasing the score of the gold triplet (v,r,w). In contrast, the $u \to v$ term captures information flow from global pseudo-negative nodes $\{u \in \mathcal{E}\}$, which serves to decrease the scores of triplets (v,r,u). Fundamentally, the term $u \to v$ is induced by the partition function in the denominator of the conditional probability (Equation 4.1). Due to the 1vsAll setting, the conditional probability $P_\theta(w \mid v,r)$ is computed over all entities in $\mathcal{E}$. As

a result, the model incorporates signals from pseudo-negative edges linking v across the entire vocabulary $\{u \in \mathcal{E}\}$ when updating the representation of the subject v. This negative contribution can be seen as a global repulsion to ensure that truly informative neighbours maintain strong influence. Note that the "negative" here is about the non-existing edges that are automatically considered due to the 1vsAll loss. This is different from the negative neighbourhood, which is from existing edges. There negative sign $\mathcal{N}_{-}^{1}[\mathrm{v}] = \{(r, s) \mid (s, r, \mathrm{v}) \in \mathcal{T}\}$ means the in-coming as opposed to outgoing.

Similarly, for the object w, we have

$$\mathrm{GD}(\phi, \{(v, r, w)\})[w] = \phi[w] + \alpha \underbrace{(1 - P_\theta(w|v, r))\, g(r) \odot \phi[v]}_{v \to w},$$

where, again, the $v \to w$ term indicates information flow from the neighbouring node v. Finally, for the nodes other than v and w, we have

$$\mathrm{GD}(\phi, \{(v, r, w)\})[u] = \phi[u] + \alpha \left(\underbrace{-P_\theta(u|v, r)\phi[v] \odot g(r)}_{v \to u} \right).$$

4.4.2 The Node View

To fully uncover the message-passing mechanism of FMs, we now focus on the gradient descent operation over a single node $v \in \mathcal{E}$, referred to as the *central node* in the GNN literature. Recalling Equation 4.8, we have:

$$\mathrm{GD}(\phi, \mathcal{T})[v] = \phi[v] + \alpha \sum_{(v,r,w) \in \mathcal{T}} \frac{\partial \log P(\mathrm{w} \mid \mathrm{v}, \mathrm{r})}{\partial \phi[v]}, \tag{4.10}$$

which aggregates the information stemming from the updates presented in the edge view. The next theorem describes how this total information flow to a particular node can be recast as an instance of message passing (cf. Section 4.3.2). We defer the full proof to Appendix B.1.1 and present a proof sketch here.

Theorem 4.4.1 (Message passing in FMs)**.** *The gradient descent operator* GD *(Equation 4.10) on the node embeddings of a DistMult model (Equation 4.4) with the maximum*

likelihood objective (Equation 4.3) and a multi-relational graph $\mathcal{T}$ defined over entities $\mathcal{E}$ induces a message-passing operator whose composing functions are:

$$q_{\mathrm{M}}(\phi[v], r, \phi[w]) = \begin{cases} \phi[w] \odot g(r) & \text{if } (r, w) \in \mathcal{N}_+^1[v], \\ (1 - P_\theta(v|w, r))\phi[w] \odot g(r) & \text{if } (r, w) \in \mathcal{N}_-^1[v]; \end{cases} \tag{4.11}$$

$$q_{\mathrm{A}}(\{m[v, r, w] \,:\, (r, w) \in \mathcal{N}^1[v]\}) = \sum_{(r,w)\in\mathcal{N}^1[v]} m[v, r, w]; \tag{4.12}$$

$$q_{\mathrm{U}}(\phi[v], z[v]) = \phi[v] + \alpha z[v] - \beta n[v], \tag{4.13}$$

where, defining the sets of triplets $\mathcal{T}^{-v} = \{(s, r, o) \in \mathcal{T} \,:\, s \neq v \wedge o \neq v\}$,

$$n[v] = \frac{|\mathcal{N}_+^1[v]|}{|\mathcal{T}|}\mathbb{E}_{P_{\mathcal{N}_+^1[v]}}\mathbb{E}_{u\sim P_\theta(\cdot|v,r)}g(r) \odot \phi[u] + \frac{|\mathcal{T}^{-v}|}{|\mathcal{T}|}\mathbb{E}_{P_{\mathcal{T}^{-v}}}P_\theta(v|s, r)g(r) \odot \phi[s], \tag{4.14}$$

where $P_{\mathcal{N}_+^1[v]}$ and $P_{\mathcal{T}^{-v}}$ are the empirical probability distributions associated to the respective sets.

Proof Sketch (Proof Sketch for Theorem 4.4.1)**.** *We outline how a single step of gradient descent (Equation 4.10) on the node embeddings of a DistMult model (Equation 4.4) with a softmax-based likelihood (Equation 4.3) induces a message-passing operator.*

Setup and Assumptions. *We consider a multi-relational graph $\mathcal{T}$ over entities $\mathcal{E}$ and relations $\mathcal{R}$. Each entity $v \in \mathcal{E}$ is associated with an embedding $\phi[v]$. The DistMult model defines the conditional probability of a tail entity given a head and relation as:*

$$P(w \mid v, r) = \frac{\exp(\Gamma(v, r, w))}{\sum_{u\in\mathcal{E}} \exp(\Gamma(v, r, u))},$$

where $\Gamma(v, r, w) = \langle\phi[v], g(r), \phi[w]\rangle$. We assume no self-loops (i.e., (v, r, v) not in $\mathcal{T}$).

Gradient Decomposition. *The gradient of the log-likelihood w.r.t. $\phi[v]$ is a sum over the triples comprising the training graph*

$$\sum_{(\mathrm{v},\mathrm{r},\mathrm{w})\in\mathcal{T}} \frac{\partial \log P(\mathrm{w} \mid \mathrm{v}, \mathrm{r})}{\partial\phi[v]},$$

which splits into:

- Outgoing edges of v: (v, r, w) *yield terms pulling $\phi[v]$ toward $\phi[\mathrm{w}] \odot g(r)$. At the*

same time, the partition function induced by the denominator of the 1vsAll loss yields terms pushing $\phi[v]$ away from global pseudo-negative entities $g(r) \odot \phi[u]$ for $u \in \mathcal{E}$ modulated by $P(u|v, r)$.

- Incoming edges of v: (w, r, v) *yield terms pulling $\phi[v]$ towards $g(r) \odot \phi[\mathrm{w}]$ modulated by $1 - P(v|\,\mathrm{w}, r)$.*

- Non-local edges*:* (s, r, o) *Triplets not involving v but in the training graph still affect $\phi[v]$ due to v's appearance in the partition function, producing a term proportional to $-P(v|s, r)g(r) \odot \phi[s]$.*

Message-Passing Form. *Collecting these categories and regrouping them based on if the term comes from v's neighbourhood, yielding:*

- *A message function $q_{\mathrm{M}}(\phi[v], r, \phi[w])$ from local neighbours, where messages along outgoing edges and incoming edges have different forms.*

- *An aggregation function q_{A} summing all messages from neighbourhood producing $z[v]$*

- *A correction term $n[v]$ from the global partition of outgoing edges and the non-local edges.*

- *An update rule:*

$$\phi[v] \leftarrow q_{\mathrm{U}}(\phi[v], z[v]) = \phi[v] + \alpha z[v] - \beta n[v],$$

with step sizes α, β.

This establishes equivalence between DistMult's gradient update and a message-passing architecture with global context.

What emerges from the equations is that each GD step contains an explicit information flow from the neighbourhood of each node, which is then aggregated with a simple summation. Through this direct information path, t GD steps cover the t-hop neighbourhood of v. As t goes towards infinity or in practice as training converges, FMs capture the global graph structure. The update function (Equation 4.13) somewhat deviates from

classic message passing frameworks as $n[v]$ of Equation 4.14 involves global information. However, we note that we can interpret this mechanism under the framework of augmented message passing [Veličković, 2022] and, in particular, as an instance of *graph rewiring*, where $n[v]$ represents rewired edges to global nodes that are not in the local neighbourhood.

Based on Theorem 4.4.1 and Equation 4.7, we can now view ϕ as the transient node states h (cf. Section 4.3.2) and GD on node embeddings as a message-passing layer. This dualism sits at the core of the ReFactor GNN model, which we describe next.

4.5 ReFactor GNN: Inductivising Factorization based Models

FMs are trained by minimising the objective (Equation 4.3), initialising both sets of parameters (ϕ and ψ) and performing GD until approximate convergence (or until early stopping terminates the training). The implications are twofold: i) the initial value of the entity lookup table ϕ does not play any major role in the final model after convergence, and ii) if we introduce a new set of entities, the conventional wisdom is to re-train[2] the model on the expanded knowledge graph. This can be computationally rather expensive and operationally complex, compared to the "inductive" models that require no additional training and can leverage node features like entity descriptions.

However, as we have just seen in Theorem 4.4.1, the training procedure of FMs may be naturally recast as a message-passing operation, which suggests that it is possible to use FMs for inductive learning tasks. In fact, we envision that there is an entire novel spectrum of model architectures interpolating between pure FMs and (various instantiations of) GNNs. Here we propose one simple implementation of such an architecture which we dub ReFactor GNNs. Figure 4.2 gives an overview of ReFactor GNNs.

The ReFactor Layer A ReFactor GNN contains L ReFactor layers, that we derive from Theorem 4.4.1. Aligning with the notations in Section 4.3.2, given a knowledge graph $\mathcal{T}$ and entity representations $h^{l-1} \in \mathbb{R}^{|\mathcal{E}| \times K}$, the ReFactor layer computes the

[2]Typically, until convergence and possibly by partially warm-starting the model parameters.

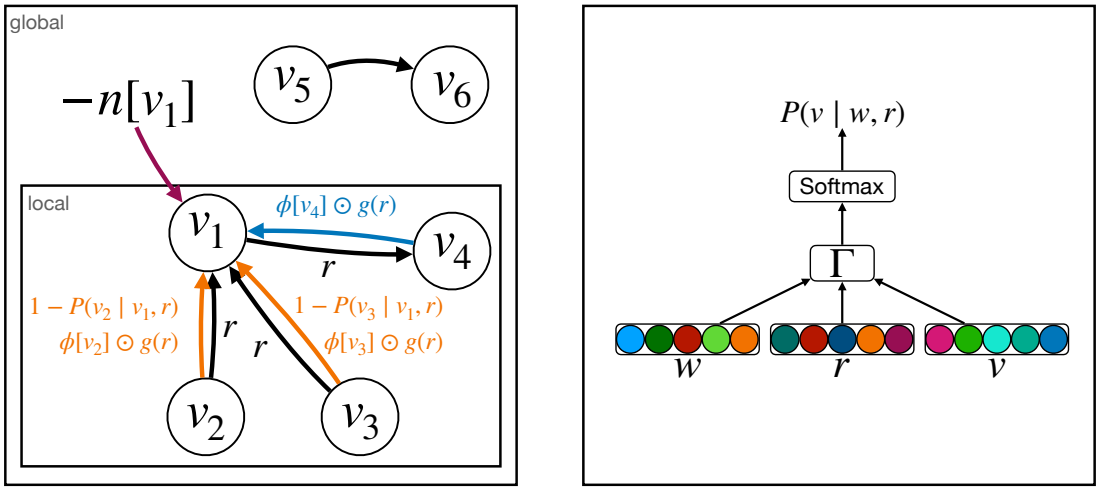

Figure 4.2: ReFactor GNN architecture. The left figure describes messages from the local neighbourhood $\{(v_2, r_1, v_1), (v_3, r_2, v_1), (v_1, r_3, v_4)\}$ (the orange and blue edges, which depend on the type of relationship of the edges) and a global normaliser term induced by the partition function (the purple arrow); The right figure describes the computation graph for calculating $P(v \mid w, r)$, where $v, w \in \mathcal{E}$ and $r \in \mathcal{R}$: the embedding representations of w, r, and v are used to score the edge (w, r, v) via the scoring function Γ, which is then normalised via the SoftMax function.

representation of a node v as follows:

$$h^l[v] = q^l(\mathcal{T}, h^{l-1})[v] = h^{l-1}[v] - \beta n^l[v] + \alpha \sum_{(r,w)\in\mathcal{N}^1[v]} q_M^l(h^{l-1}[v], r, h^{l-1}[w]), \quad (4.15)$$

where the terms n^l and q_{M}^l are derived from Equation 4.14 and Equation 4.11, respectively. We note that ReFactor GNNs treat incoming and outgoing neighbourhoods differently instead of treating them equally as in for example the R-GCN, the first GNN on multi-relational graphs [Schlichtkrull et al., 2018].

Equation 4.15 describes the full batch setting, which can be expensive if the KG contains many edges. Therefore, in practice, whenever the graph is big, we adopt a stochastic evaluation of the ReFactor layer by decomposing the evaluation into several mini-batches. We partition $\mathcal{T}$ into a set of computationally tractable mini-batches. **For each mini-batch, we restrict the neighbourhoods to the subgraph induced by it and readjust the computation of $n^l[v]$ to include only entities and edges present in it**.

We leave the investigation of other stochastic strategies (e.g. by taking Monte Carlo estimations of the expectations in Equation 4.14) to future work. Finally, we cascade the mini-batch evaluation to produce one full layer evaluation (i.e. one message-passing round over the entire graph).

Training The learnable parameters of ReFactor GNNs are the relation embeddings ψ, which parameterise the $g(r)$ in the message function $q_M^l, l \in [1, L]$. Inspired by Fey et al. [2021], You et al. [2020], we learn ψ by layer-wise (stochastic) gradient descent. This is in contrast to conventional GNN training, where one needs to backpropagate through all the message-passing layers $l \in [1, L]$. A (full-batch) GD training dynamic for ψ can be written as

$$\psi_{t+1} = \psi_t - \eta \nabla \mathcal{L}_t(\psi_t)$$

with:

$$\mathcal{L}_t(\psi_t) = \sum_{\mathcal{T}} -\frac{1}{|\mathcal{T}|} \log P_{\psi_t}(w|v, r)$$

$$\text{where} \quad P_{\psi_t}(w|v, r) = \text{Softmax}(\Gamma(v, r, \cdot))[w] \quad \Gamma(v, r, w) = \langle h^t[v], h^t[w], g_{\psi_t}(r) \rangle.$$

$h^t[\cdot]$ denotes the node state of a particular node at iteration t and the node state is updated recursively as

$$\begin{aligned} h^0 &= X, \text{initial node features} \\ h^t &= q^l(\mathcal{T}, h^{t-1}) \text{where } l = t \bmod L, t \geq 1. \end{aligned} \tag{4.16}$$

This dynamic ensures that at each step t, only the current layer $l = t \bmod L \in [1, L]$ is activated and participates in the backpropagation. Early layers $< l$ are truncated from the computational graph by treating h^{t-1} as a fixed (non-differentiable) input for the current layer, bounding the gradient path to a single layer per training step. Such truncation of the computational graph to reduce memory usage is not uncommon and have been used in meta-learning algorithms [Chen et al., 2019] and for GNN scaling techniques [Fey et al., 2021].

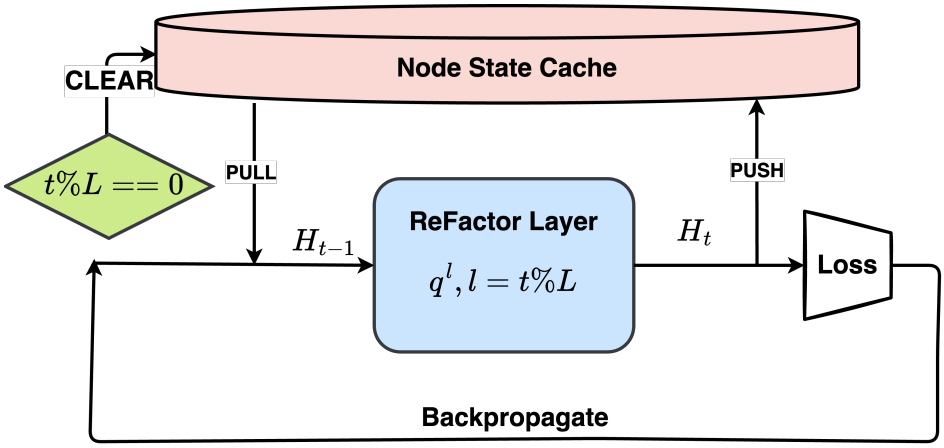

Figure 4.3: Illustration of the external node state cache used during training.

External cache, its push, pull, and clear. Implementation-wise, such a training dynamic equals to maintaining an external **memory** for storing and retriving historical node states h^{t-1} to compute h^t using Equation 4.15. Figure 4.3 illustrates the external cache. During the model optimisation, the historical node states are fixed. After each training step, newly computed node states are pushed to update the historical cache. But this push occurs after gradient computation, and these historical vectors are not part of the current backpropagation path. After every L full batches, we clear the cache by resetting all node states in the cache to their initial input values X (e.g., textual or random features). This procedure of push, pull, and clear, emulates an unrolling of the message-passing dynamic up to L layers, and forces the model to predict based on on-the-fly L-layer message-passing. After training, we obtain ψ^* and perform inference by running L-layer message-passing with ψ^*. In general, L determines the number of effective message-passing layers in ReFactor GNNs. A larger L enables ReFactor GNNs to fuse information from more hops of direct neighbourhoods into the final node representations. In the meantime, it reduces the inductive applicability of ReFactor GNNs due to over-smoothing and computational requirements. In the extreme case of $L = \infty$, *where we never clear the node state cache during training*, the final cached node states will be used for inference. Note that this latter inference regime is inherently transductive since there will be no cached states for new nodes. Future work may explore a more streamlined implementation by simply resetting the entity embeddings periodically as in Chen et al. [2023].

Relation to prior work While our use of caching is inspired by AutoScale [Fey et al., 2021], our model diverges in key ways. Unlike Fey et al. [2021], where the historical node states are only used for out-of-batch neighbour nodes, we use historical node states for all nodes. Fey et al. [2021] define only the "push" and "pull" operations for the memory. We additionally define a "clear" operation for the memory. This cache-clearing mechanism acts as a form of *active forgetting*, which we introduce to promote inductive capability. Work in the spirit of active forgetting has been extensively explored in the continual learning literature as a mechanism for improving adaptability and reduce overfitting to past learnings. For instance, neural pruning removes low-activity neurons to free capacity for future tasks [Golkar et al.]; episodic backward updates selectively discard outdated gradients to favor recent learning [Lee et al., 2019]; and meta-experience replay strategies [Riemer et al.] reduce gradient interference, effectively suppressing conflicting knowledge. Our cache reset parallels these approaches by clearing outdated node embeddings, thereby preventing over-specialization and supporting generalization to unseen entities. These modifications are essential in adapting static factorisation models into a dynamic, message-passing framework suitable for both transductive and inductive link prediction tasks.

4.6 Experiments

We perform experiments to answer the following questions regarding ReFactor GNNs:

- **RQ1.** ReFactor GNNs are derived from a message-passing reformulation of FMs: do they also inherit FMs' predictive accuracy in *transductive* KGC tasks? (Section 4.6.1)
- **RQ2.** ReFactor GNNs "inductivise" FMs. Are they more statistically accurate than other GNN baselines in *inductive* KGC tasks? (Section 4.6.2)
- **RQ3.** The term $n[v]$ involves nodes that are not in the 1-hop neighbourhood. Is such *augmented message passing* [Veličković, 2022] necessary for good KGC performance? (Section 4.6.3)

For transductive experiments, we used three well-established KGC datasets: *UMLS*, *CoDEx-S*, and *FB15K237* [Kemp et al., 2006, Safavi and Koutra, 2020, Toutanova and Chen, 2015]. For inductive experiments, we used the inductive KGC benchmarks introduced by GraIL [Teru et al., 2020], which include 12 pairs of knowledge graphs:

- (*FB15K237_vi*, *FB15K237_vi_ind*),
- (*WN18RR_vi*, *WN18RR_vi_ind*),
- (*NELL_vi*, *NELL_vi_ind*),

where $i \in [1, 2, 3, 4]$, and (*_vi*, *_vi_ind*) represents a pair of graphs with **a shared relation vocabulary and non-overlapping entities**. Note that the GraIL setup is different from a completely inductive setup, where both the relations and entities are unseen at test time.

We follow the standard KGC evaluation protocol by fully ranking all the candidate entities and computing two metrics using the ranks of the ground-truth entities: Mean Reciprocal Ranking (MRR) and Hit Ratios at Top K (Hits@K) with $K \in [1, 3, 10]$. For the inductive KGC, we additionally consider the partial-ranking evaluation protocol used by GraIL for a fair comparison. Empirically, we find full ranking more difficult than partial ranking, and thus more suitable for reflecting the differences among models on GraIL datasets. In fact, we would like to call for future work on GraIL datasets to also adopt a full ranking protocol on these datasets.

Our *transductive* experiments used $L = \infty$, i.e. node states cache is never cleared, as we wanted to see if ReFactor GNNs ($L = \infty$) can reach the performance of the FMs (Section 4.6.1); on the other hand, in our *inductive* experiments, we used ReFactor GNNs with $L \in \{1, 2, 3, 6, 9\}$, since we wanted to test their performances in inductive settings akin to standard GNNs (Section 4.6.2). We used a hidden size of 768 for the node representations. All the models are trained using $[128, 512]$ in-batch negative samples and one global negative node for each positive link. We performed a grid search over the other hyperparameters and selected the best configuration based on the validation MRR. Since training deep GNNs with full-graph message passing might be slow for large knowledge graphs, we follow the literature [Hamilton et al., 2017, Zou et al., 2019, Zeng et al., 2020] to sample sub-graphs for training GNNs as indicated by Equation 4.6. Considering that sampling on the fly often prevents high

Table 4.1: Test MRR for transductive KGC tasks.

Entity Encoder	UMLS	CoDEx-S	FB15K237
R-GCN	–	0.33	0.25
Lookup (FM, specif. DistMult)	0.90	0.43	0.30
ReFactor GNNs ($L = \infty$)	0.93	0.44	0.33

utilisation of GPUs, we resort to a two-stage process: we first sampled and serialised sub-graphs around the target edges in the mini-batches; we then trained the GNNs with the serialised sub-graphs. To ensure that we have sufficient sub-graphs for training the models, we sampled for 20 epochs for each knowledge graph, i.e. 20 full passes over the full graph. The sub-graph sampler we currently used is LADIES [Zou et al., 2019].

4.6.1 RQ1: ReFactor GNNs for Transductive Learning

ReFactor GNNs are derived from the message-passing reformulation of FMs. We expect them to approximate the performance of FMs for transductive KGC tasks. To verify this, we perform experiments on the datasets UMLS, CoDEx-S, and FB15K237. For a fair comparison, we use Equation 4.2 as the decoder and consider i) lookup embedding table as the entity encoder, which forms the FM when combined with the decoder (Section 4.3.1), and ii) ReFactor GNNs as the entity encoder. Note that the equivalence between ReFactor GNNs and the standard FMs are only obtained when ReFactor GNNs are trained with $L = \infty$, i.e. we never clear the node state cache. This is different from inductive setups, where ReFactor GNNs are trained with a finite L. Since transductive KGC tasks do not involve new entities, the node state cache in ReFactor GNNs can be directly used for link prediction. Table 4.1 summarises the result. We observe that ReFactor GNNs achieve a similar or slightly better performance compared to the FM. This shows that ReFactor GNNs are able to capture the essence of FMs and thus remain competitive at transductive KGC.

4.6.2 RQ2: ReFactor GNNs for Inductive Learning

Despite FMs' good empirical performance on transductive KGC tasks, they fail to be as inductive as GNNs. According to our reformulation, this is due to the infinite message-

passing layers hidden in FMs' optimisation. Discarding infinite message-passing layers, ReFactor GNNs enable FMs to perform inductive reasoning tasks by learning to use a finite set of message-passing layers for prediction similarly to GNNs.

Here we present experiments to verify ReFactor GNNs's capability for inductive reasoning. Specifically, we study the task of inductive KGC and investigate whether ReFactor GNNs can generalise to unseen entities. Following Teru et al. [2020], on GraIL datasets, we trained models on the original graph, and run 0-shot link prediction on the *_ind* test graph. Similar to the transductive experiments, we use Equation 4.2 as the decoder and vary the entity encoder. We denote three-layer ReFactor GNNs as `ReFactor(3)` and six-layer ReFactor GNNs as `ReFactor(6)`. We consider several baseline entity encoders: i) `no-pretrain`, models without any pretraining on the original graph; ii) `GAT(3)`, three-layer graph attention network [Veličković et al., 2018]; iii) *GAT(6)*, six-layer graph attention network; iv) `GraIL`, a sub-graph-based relational GNN [Teru et al., 2020]; v) `NBFNet`, a path-based GNN [Zhu et al., 2021], current SoTA on GraIL datasets. In addition to randomly initialised vectors as the node features, we also used textual node features, RoBERTa [Liu et al., 2019a] Encodings of the entity descriptions, which are produced by SentenceBERT [Reimers and Gurevych, 2019]. Due to space reason, we present the results on (*FB15K237_v*1, *FB15K237_v*1*_ind*) in Figure 4.4. Results on other datasets are similar and can be found in the appendix. We can see that without textual node features, ReFactor GNNs perform better than GraIL (+23%); with textual node features, ReFactor GNNs outperform both GraIL (+43%) and NBFNet (+10%), achieving new SoTA results on inductive KGC.

Performance vs Parameter Efficiency as #Message-Passing Layers Increases Usually, as the number of message-passing layers increases in GNNs, the over-smoothing issue occurs while the computational cost also increases exponentially. ReFactor GNNs avoid this by layer-wise training and sharing the weights across layers. Here we compare ReFactor GNNs with $\{1, 3, 6, 9\}$ message-passing layer(s) with same-depth GATs. Results are summarised in Figure 4.5. We observe that increasing the number of message-passing layers in GATs does not necessarily improve the predictive accuracy – the best results were obtained with 3 message-passing layers on *FB15K237_v1* while using 6 and 9 layers leads to performance degradation. On the other hand, ReFactor GNNs obtain consistent improvements when increasing #Layers from 1 to 3, 6, and 9. ReFactor

Figure 4.4: Inductive KGC performance, trained on the KG *FB15K237_v1* and tested on another KG *FB15K237_v1_ind*, where the entities are completely new. The results of GraIL and NBFNet are taken from Zhu et al. [2021]. The grey bars indicate methods that are not devised to incorporate node features.

GNNs $(6, 6)$ and $(9, 9)$ clearly outperform their GAT counterparts. Most importantly, ReFactor GNNs are more parameter-efficient than GATs, with a constant #Parameters as #Layers increases.

4.6.3 RQ3: Beyond Message-Passing

As shown by Theorem 4.4.1, ReFactor GNNs contain not only terms capturing information flow from the 1-hop neighbourhood, which falls into the classic message-passing framework, but also a term $n[v]$ that involve nodes outside the 1-hop neighbourhood. The term $n[v]$ can be treated as *augmented message-passing* on a dynamically rewired graph [Veličković, 2022]. Here we perform ablation experiments to measure the impact of the $n[v]$ term. Table 4.2 summarises the ablation results: we can see that, without the term $n[v]$, ReFactor GNNs with random vectors as node features yield a 2% lower MRR, while ReFactor GNNs with RoBERTa textual encodings as node features

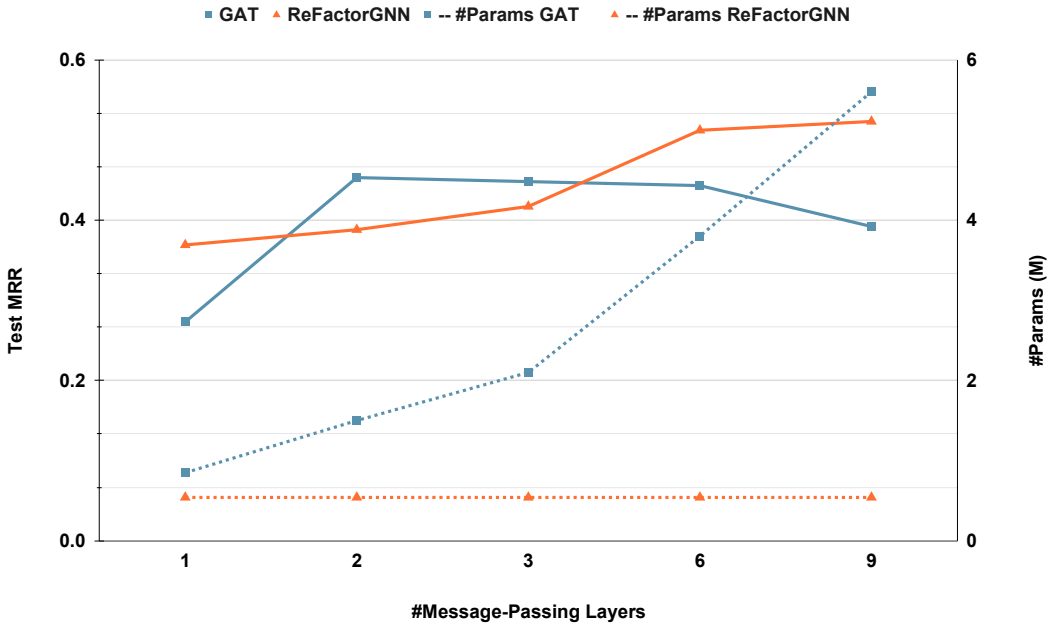

Figure 4.5: Performance vs parameter efficiency on *FB15K237_v1*. Left axis is Test MRR while right axis is #parameters. The solid lines and dashed lines indicate the changes of Test MRR and the changes of #parameters.

Table 4.2: Ablation on $n[v]$ for ReFactor GNNs (6) trained on *FB15K237_v1*.

Test MRR	With Random Features	With Textual Features
with $n[v]$	0.425	0.486
without $n[v]$	0.418	0.452

produce a 7% lower MRR. This suggests that augmented message-passing also plays a significant role in ReFactor GNNs' generalisation properties in downstream link prediction tasks. Future work might gain more insights by further dissecting the $n[v]$ term.

4.7 Discussion

Summary. The task of multi-relational link prediction forms the cornerstone of constructing useful knowledge graphs, which, in turn, underpin modern knowledge engines. Factorization Models (FMs) and Graph Neural Networks (GNNs) are two prominent ap-

proaches for this task. FMs excel in transductive settings, while GNNs are better suited for inductive scenarios. Despite the sharp differences in their analytical forms, our work establishes a link between FMs and GNNs. By reformulating FMs as GNNs, we address a critical question: why are FMs superior transductive multi-relational link predictors but fail in inductive scenarios? The answer lies in FMs performing excessive message-passing in standard training, capturing excessive global structures, and producing overly rigid representations.

Building on this insight, we propose ReFactor GNNs, a novel GNN variant that incorporates an *active forgetting* mechanism into the message-passing process of FMs. ReFactor GNNs periodically reset the cache of prior message-passing computations, enabling the model to focus on local neighbourhood information instead of over-relying on the entire training graph. Empirical experiments demonstrate that ReFactor GNNs achieve significantly higher accuracy than GNN baselines on inductive link prediction tasks, bridging the gap between the strengths of FMs and GNNs.

Limitations. Since we adopted a two-stage (sub-graph serialisation and then model training) approach instead of online sampling, there can be side effects from the low sub-graph diversity. In our experiments, we used LADIES [Zou et al., 2019] for sub-graph sampling. Experiments with different sub-graph sampling algorithms, such as GraphSaint [Zeng et al., 2020] might affect the downstream link prediction results. Furthermore, it would be interesting to analyse decoders other than DistMult, as well as additional optimisation schemes beyond SGD and AdaGrad. We do not dive deeper into the expressiveness of ReFactor GNNs. Nevertheless, we offer a brief discussion in Section B.1.1.

Implications. The most direct future work would be using the insight to develop more sophisticated models at the intersection between FMs and GNNs, e.g. by further parameterising the message/update function. One implication from our work is that reformulating FMs as message-passing enables the idea of "learning to factorize". This might broaden the usage of FMs, going beyond link prediction, to tasks such as graph classification. Another implication comes from our approach of unpacking embedding updates into a series of message-passing operations. This approach can be generalised to other dot-product-based models that use embedding layers for processing the inputs, lend-

ing it naturally to understanding complicated attention-based models like Transformers. Although Transformers can be treated as GNNs over fully-connected graphs, where a sentence would be a graph and its tokens would be the nodes, the message-passing is limited to within each sentence under this view. We instead envision cross-sentence message-passing by reformulating the updates of the token embedding layer in transformers. In general, the direction of organising FMs, GNNs, and transformers under the same framework will allow a better understanding of all three models. While FMs and GNNs excel in the structured paradigm, transformers are often the default choice for the unstructured paradigm. Unveiling the connections among these models can facilitate the seamless integration of the structured and unstructured paradigm, paving the way for building universal knowledge engines.

Bibliography

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O Alabi, Shamsuddeen H Muhammad, Peter Nabende, et al. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4488–4508. Association for Computational Linguistics (ACL), 2022.

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne R. Tamang, and Robert Rallo. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics. *ArXiv*, abs/1907.08650, 2019. URL `https://api.semanticscholar.org/CorpusID:198147334`.

Divyanshu Aggarwal, Ashutosh Sathe, and Sunayana Sitaram. Exploring pretraining via active forgetting for improving cross lingual transfer for decoder language models. *arXiv preprint arXiv:2410.16168*, 2024.

Jethro Akroyd, Sebastian Mosbach, Amit Bhave, and Markus Kraft. Universal digital twin - a dynamic knowledge graph. *Data-Centric Engineering*, 2:e14, 2021. doi: 10.1017/dce.2021.10.

Ibrahim Alabdulmohsin, Hartmut Maennel, and Daniel Keysers. The impact of

reinitialization on generalization in convolutional neural networks. *arXiv preprint arXiv:2109.00267*, 2021.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Prithviraj Ammanabrolu and Mark Riedl. Learning knowledge graph-based world models of textual environments. *Advances in Neural Information Processing Systems*, 34: 3720–3731, 2021.

Michael C. Anderson and Justin C. Hulbert. Active forgetting: Adaptation of memory by prefrontal control. *Annual Review of Psychology*, 72(1):1–36, 2021. doi: 10.1146/annurev-psych-072720-094140. URL https://doi.org/10.1146/annurev-psych-072720-094140. PMID: 32928060.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2020.

Various authors. Wikipedia, the free encyclopedia, 2024. URL https://www.wikipedia.org. A collaboratively edited, free online encyclopedia.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Tucker: Tensor factorization for knowledge graph completion. In *EMNLP/IJCNLP*, 2019.

Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

Jeffrey Barrett and Kevin JS Zollman. The role of forgetting in the evolution and learning of language. *Journal of Experimental & Theoretical Artificial Intelligence*, 21(4):293–309, 2009.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. *arXiv preprint arXiv:2002.09571*, 2020.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Edward L Bennett, Marian C Diamond, David Krech, and Mark R Rosenzweig. Chemical and anatomical plasticity of brain: Changes in brain through experience, demanded by learning theories, are found in experiments with rats. *Science*, 146(3644):610–619, 1964.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8(3):3–24, 2007.

Jacob A Berry, Dana C Guhle, and Ronald L Davis. Active forgetting and neuropsychiatric diseases. *Molecular Psychiatry*, pages 1–11, 2024.

Tarek R Besold, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, et al. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pages 1–51. IOS press, 2021.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, J. Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

Léon Bottou. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_25. URL `https://doi.org/10.1007/978-3-642-35289-8_25`.

Thorsten Brants, Ashok Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference*

on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 858–867, 2007.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Richard E Brown. Hebb and cattell: The genesis of the theory of fluid and crystallized intelligence. *Frontiers in human neuroscience*, 10:606, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Jerome Bruner. *The Process of Education*. Harvard University Press, 1960.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. Imbalanced learning: A comprehensive evaluation of resampling methods for class imbalance. *arXiv preprint arXiv:1710.05381*, 2018. URL `https://arxiv.org/abs/1710.05381`.

Raymond B Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1, 1963.

Yihong Chen, Bei Chen, Xiangnan He, Chen Gao, Yong Li, Jian-Guang Lou, and Yue Wang. λopt: Learn to regularize recommender models in finer levels. In *KDD 2019 (Oral), Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 978–986, 2019.

Yihong Chen, Pasquale Minervini, Sebastian Riedel, and Pontus Stenetorp. Relation prediction as an auxiliary training objective for improving multi-relational graph representations. In *AKBC 2021*, 2021.

Yihong Chen, Pushkar Mishra, Luca Franceschi, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Refactor gnns: Revisiting factorisation-based models from a message-passing perspective. In *Advances in Neural Information Processing Systems*, 2022.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. Improving language plasticity via pretraining with active forgetting. In *NeurIPS 2023*, 2023.

Yihong Chen, Xiangxiang Xu, Yao Lu, Pontus Stenetorp, and Luca Franceschi. Jet expansions of residual computation, 2024. URL `https://arxiv.org/abs/2410.06024`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Brandon C Colelough and William Regli. Neuro-symbolic ai in 2024: A systematic review. 2024.

Together Computer. Redpajama dataset. `https://www.together.xyz/blog/redpajama`, 2023. Accessed: 2023-12-12.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.,

2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. Investigating critical period effects in language acquisition through neural language models. *arXiv preprint arXiv:2407.19325*, 2024.

OpenWebText Contributors. The openwebtext dataset. https://github.com/jcpeterson/openwebtext, 2019. Accessed: 2023-12-12.

Moheb Costandi. *Neuroplasticity*. MIT Press, 2016.

Common Crawl. Common crawl corpus. https://commoncrawl.org, 2023. Accessed: 2023-12-12.

Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

Gilles Deleuze and Paul Patton. *Difference and Repetition*. Athlone, London, 1994.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Kevin P. Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279–90, 2006. URL `https://api.semanticscholar.org/CorpusID:22491040`.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=OpC-9aBBVJe`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

David Steven Dummit, Richard M Foote, et al. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, 2022.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma,

Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. *arXiv preprint arXiv:2403.00824*, 2024.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Jure Leskovec. Gnnautoscale: Scalable and expressive graph neural networks via historical embeddings. In *ICML*, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

Eberhard Fuchs and Gabriele Flügge. Adult neuroplasticity: more than 40 years of research. *Neural plasticity*, 2014(1):541870, 2014.

A Garcez, M Gori, LC Lamb, L Serafini, M Spranger, and SN Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4):611–632, 2019.

Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UKr0MwZM6fL.

Floris Geerts and Juan L Reutter. Expressiveness and approximation properties of graph neural networks. In *International Conference on Learning Representations*, 2021.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. URL https://arxiv.org/abs/2004.07780.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Siavash Golkar, Micheal Kagan, and Kyunghyun Cho. Continual learning via neural pruning. In *Real Neurons {\&} Hidden Units: Future directions at the intersection of neuroscience and artificial intelligence@ NeurIPS 2019*.

Joshua T Goodman. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434, 2001.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 2:729–734 vol. 2, 2005.

C Shawn Green and Daphne Bavelier. Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychology and aging*, 23(4):692, 2008.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

Axel Guskjolen and Mark S Cembrowski. Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular psychiatry*, 28(8):3207–3219, 2023.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 15(4):1265–1306, 2024.

William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Oliver Hardt, Einar Örn Einarsson, and Karim Nader. A bridge over troubled water: Reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual review of psychology*, 61(1):141–167, 2010.

Oliver Hardt, Karim Nader, and Lynn Nadel. Decay happens: the role of active forgetting in memory. *Trends in cognitive sciences*, 17(3):111–120, 2013.

Michael Hart and Project Gutenberg Volunteers. Project gutenberg online library, 1971–2024. URL `https://www.gutenberg.org`. Free eBooks from the public domain.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL `https://aclanthology.org/2022.acl-long.234`.

Felix Hausdorff. *Set theory*, volume 119. American Mathematical Soc., 2021.

Frederick Hayes-Roth, Donald A Waterman, and Douglas B Lenat. *Building expert systems*. Addison-Wesley Longman Publishing Co., Inc., 1983.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021a.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=XPZIaotutsD.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys*, 6(1):164–189, 1927.

S Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation MIT-Press*, 1997.

Sara Hooker. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

Roy Horan. The neuropsychological connection between creativity and meditation. *Creativity research journal*, 21(2-3):199–222, 2009.

John L Horn and Raymond B Cattell. Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, 57(5):253, 1966.

Ian Horrocks. Owl: A description logic based ontology language. In *International conference on principles and practice of constraint programming*, pages 5–8. Springer, 2005.

Ian Horrocks, Peter F Patel-Schneider, and Frank van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003. URL https://doi.org/10.1016/j.websem.2003.07.001.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

David Hume. An enquiry concerning human understanding. 1748. *Classics of Western Philosophy*, pages 763–828, 1999.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

Alex Iacob, Lorenzo Sani, Meghdad Kurmanji, William F Shen, Xinchi Qiu, Dongqi Cai, Yan Gao, and Nicholas D Lane. Dept: Decoupled embeddings for pre-training language models. *arXiv preprint arXiv:2410.05021*, 2024.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Qun8fv4qSby`.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6t0Kwf8-jrj`.

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Knowledge base completion: Baseline strikes back (again). *ArXiv*, abs/2005.00804, 2020a.

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Knowledge base completion: Baseline strikes back (again). *CoRR*, abs/2005.00804, 2020b. URL `https://arxiv.org/abs/2005.00804`.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. A survey on knowledge graphs: Representation, acquisition and applications. *arXiv preprint arXiv:2002.00388*, 2020.

Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. Knowledge base completion: Baselines strike back. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay B. Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 69–74. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-2609. URL `https://doi.org/10.18653/v1/w17-2609`.

Immanuel Kant. *Critique of Pure Reason (1st edition)*. Macmillan Company, Mineola, New York, 1781.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

Jill L Kays, Robin A Hurley, and Katherine H Taber. The dynamic brain: neuroplasticity and mental health. *The Journal of neuropsychiatry and clinical neurosciences*, 24(2): 118–124, 2012.

Seyed Mehran Kazemi and David Poole. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31, 2018.

Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, pages 381–388. AAAI Press, 2006.

Phillip Kent. Fluid intelligence: A brief history. *Applied Neuropsychology: Child*, 6(3): 193–203, 2017.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Byung-Hak Kim, Arvind Yedla, and Henry D Pfister. Imp: A message-passing algorithm for matrix completion. In *2010 6th International Symposium on Turbo Codes & Iterative Information Processing*, pages 462–466. IEEE, 2010.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jeffrey A Kleim and Theresa A Jones. Principles of experience-dependent neural plasticity: implications for rehabilitation after brain damage. 2008.

Donald Ervin Knuth. *The art of computer programming*, volume 3. Pearson Education, 1997.

Stanley Kok and Pedro M. Domingos. Statistical predicate invention. In *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 433–440. ACM, 2007.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2869–2878. PMLR, 2018.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Trans. Intell. Syst. Technol.*, January 2025. ISSN 2157-6904. doi: 10.1145/3712701. URL https://doi.org/10.1145/3712701. Just Accepted.

Su Young Lee, Choi Sungik, and Sae-Young Chung. Sample-efficient deep reinforcement learning via episodic backward update. *Advances in neural information processing systems*, 32, 2019.

Benedetta Leuner and Elizabeth Gould. Structural plasticity and hippocampal function. *Annual review of psychology*, 61(1):111–140, 2010.

Benjamin J Levy, Nathan D McVeigh, Alejandra Marful, and Michael C Anderson. Inhibiting your native language: The role of retrieval-induced forgetting during second-language acquisition. *Psychological Science*, 18(1):29–34, 2007.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, 2020a.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020b.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.143. URL https://aclanthology.org/2021.acl-long.143.

Chenchen Li, Aiping Li, Ye Wang, Hongkui Tu, and Yichen Song. A survey on approaches and applications of knowledge representation learning. In *2020 IEEE Fifth*

International Conference on Data Science in Cyberspace (DSC), pages 312–319. IEEE, 2020.

Ren Li, Yanan Cao, Qiannan Zhu, Guanqun Bi, Fang Fang, Yi Liu, and Qian Li. How does knowledge graph embedding extrapolate to unseen data: a semantic evidence view. *CoRR*, abs/2109.11800, 2021b.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. Structure-aware language model pretraining improves dense retrieval on structured data. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In *ICLR (Poster)*, 2016.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR, 2022.

Chen Liu, Jonas Pfeiffer, Anna Korhonen, Ivan Vulić, and Iryna Gurevych. Delving deeper into cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2408–2423, 2023a.

Chunan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. Asep: Benchmarking deep learning methods for antibody-specific epitope prediction. In *NeurIPS 2024, Proceedings of the Thirty-eighth Conference on Neural Information Processing Systems, Datasets and Benchmarks*, 2024a.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024b.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019b. URL `http://arxiv.org/abs/1907.11692`.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities, 2024. URL `https://arxiv.org/abs/2408.04682`.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '21, page 1150–1160, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467350. URL `https://doi.org/10.1145/3447548.3467350`.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett,

editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23190–23211. PMLR, 7 2023. URL https://proceedings.mlr.press/v202/lyle23b.html.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Brian MacWhinney. A unified model of language acquisition. In Judith F. Kroll and Annette M.B. De Groot, editors, *Handbook of Bilingualism: Psycholinguistic Approaches*, pages 49–67. Oxford University Press, 2005.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *ACL 2023, Findings of the Association for Computational Linguistics*, 2023.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022.

B. D. Mishra, Niket Tandon, and P. Clark. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246, 2017.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2021.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.

Sameh K Mohamed, Vít Nováček, Pierre-Yves Vandenbussche, and Emir Muñoz. Loss functions in knowledge graph embedding models. In *Proceedings of DL4KG2019-Workshop on Deep Learning for Knowledge Graphs*, page 1, 2019.

Aaron Mueller. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. *arXiv preprint arXiv:2407.04690*, 2024.

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4710–4723, 2019.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *NAACL-HLT (2)*, pages 327–333. Association for Computational Linguistics, 2018.

Timothy Nguyen. Understanding transformers via n-gram statistics. *arXiv preprint arXiv:2407.12034*, 2024.

M. Nickel, Volker Tresp, and H. Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, 2011a.

M. Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104:11–33, 2016a.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, pages 809–816. Omnipress, 2011b.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proc. IEEE*, 104(1):11–33, 2016b.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In *AAAI*, pages 1955–1961. AAAI Press, 2016c.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International Conference on Machine Learning*, pages 16828–16847. PMLR, 2022.

Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and Andre Barreto. Deep reinforcement learning with plasticity injection. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. URL https://openreview.net/forum?id=O9cJADBZT1.

nostalgebraist. interpreting gpt: the logit lens, 2021. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens#HEf5abD7hqqAY2GSQ.

Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):48–75, 2019.

Simon Nørby. Why forget? on the adaptive value of memory loss. *Perspectives on Psychological Science*, 10(5):551–578, 2015. doi: 10.1177/1745691615596787. URL https://doi.org/10.1177/1745691615596787. PMID: 26385996.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113: 54–71, 2019.

Denise C Park and Chih-Mao Huang. Culture wires the brain: A cognitive neuroscience perspective. *Perspectives on Psychological Science*, 5(4):391–400, 2010.

Bernhard Pastötter, Karl-Heinz Bäuml, and Simon Hanslmayr. Oscillatory brain activity before and after an internal context change—evidence for a reset of encoding processes. *NeuroImage*, 43(1):173–181, 2008.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Judea Pearl and Glenn Shafer. Probabilistic reasoning in intelligent systems: Networks of plausible inference. *Synthese-Dordrecht*, 104(1):161, 1995.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, 2019.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL `https://aclanthology.org/2020.emnlp-main.617`.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Unks everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, 2021.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, 2022.

Jean Piaget. *The Child's Conception of the World*. Harcourt, Brace & World, 1929.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.

BigCode Project. Starcoder dataset. `https://huggingface.co/bigcode`, 2023. Accessed: 2023-12-12.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

Vijaya Raghavan T Ramkumar, Elahe Arani, and Bahram Zonooz. Learn, unlearn and relearn: An online learning paradigm for deep neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=WN1O2MJDST`.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Siamak Ravanbakhsh, Barnabás Póczos, and Russell Greiner. Boolean matrix factorization and noisy completion via message passing. In *International Conference on Machine Learning*, pages 945–954. PMLR, 2016.

Michael Reed and Barry Simon. *Methods of modern mathematical physics: Functional analysis*, volume 1. Gulf Professional Publishing, 1980.

Benjamin Reichman and Larry Heck. Dense passage retrieval: Is it retrieving?, 2024. URL https://arxiv.org/abs/2402.11035.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, 2020.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mark R Rosenzweig. Aspects of the search for neural mechanisms of memory. *Annual review of psychology*, 47(1):1–32, 1996.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You can teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkxSmlBFvr.

Tomás J Ryan and Paul W Frankland. Forgetting as a form of adaptive engram cell plasticity. *Nature Reviews Neuroscience*, 23(3):173–186, 2022.

Tara Safavi and Danai Koutra. Codex: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8328–8350, 2020.

Rolf Sandell. Structural change and its assessment. *International Journal of Psychology and Psychoanalysis*, 5:042, 2019. doi: 10.23937/2572-4037.1510042.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1): 61–80, 2009.

Tom Schaul and Jürgen Schmidhuber. Metalearning. *Scholarpedia*, 5(6):4650, 2010.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.

Hans-Jörg Schmid. *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*. American Psychological Association, 2017.

Jürgen Schmidhuber. Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Frontiers in psychology*, 4:313, 2013.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. URL `https://arxiv.org/abs/2102.11107`.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR, 2018.

Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and editing predictions by modeling model computation. *arXiv preprint arXiv:2404.11534*, 2024.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Yifei Shen, Yongji Wu, Yao Zhang, Caihua Shan, Jun Zhang, Khaled B Letaief, and Dongsheng Li. How powerful is graph convolution for recommendation? *arXiv preprint arXiv:2108.07567*, 2021.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Herbert A Simon et al. Invariants of human behavior. *Annual review of psychology*, 41 (1):1–20, 1990.

Burrhus Frederic Skinner. *Science and human behavior*. Number 92904. Simon and Schuster, 1965.

Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, `https://github.com/allenai/pes2o`.

Balasubramaniam Srinivasan and Bruno Ribeiro. On the equivalence between positional node embeddings and structural graph representations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJxzFySKwH`.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5516–5522, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.489. URL https://aclanthology.org/2020.acl-main.489.

Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha P. Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5516–5522. Association for Computational Linguistics, 2020b. URL https://www.aclweb.org/anthology/2020.acl-main.489/.

Anej Svete and Ryan Cotterell. Transformers can represent n-gram language models. *arXiv preprint arXiv:2404.14994*, 2024.

Ahmed Taha, Abhinav Shrivastava, and Larry S Davis. Knowledge evolution in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12843–12852, 2021.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401, 2020. URL https://api.semanticscholar.org/CorpusID:220936592.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Komal K. Teru, Etienne G. Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9448–9457. PMLR, 2020.

Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. Limits of transformer language models on learning to compose algorithms. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. URL `https://arxiv.org/abs/2402.05785`.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Susumu Tonegawa, Xu Liu, Steve Ramirez, and Roger Redondo. Memory engram cells have come of age. *Neuron*, 87(5):918–931, 2015.

Susumu Tonegawa, Mark D Morrissey, and Takashi Kitamura. The role of engram cells in the systems consolidation of memory. *Nature Reviews Neuroscience*, 19(8):485–498, 2018.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 57–66, 2015.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1174. URL `https://www.aclweb.org/anthology/D15-1174`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA, 6 2016. PMLR. URL https://proceedings.mlr.press/v48/trouillon16.html.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BylA_C4tPr.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Petar Veličković. Message passing all the way up, 2022. URL https://arxiv.org/abs/2202.11097.

Tom Veniat, Ludovic Denoyer, and Marc'Aurelio Ranzato. Efficient continual learning with modular networks and task-driven priors. *arXiv preprint arXiv:2012.12631*, 2020.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 1288–1301, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.75.

Lev Vygotsky. *Thought and Language*. MIT Press, 1934.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. On the exploitability of reinforcement learning with human feedback for large language models. *arXiv preprint arXiv:2311.09641*, 2023.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2309.05660`.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021. URL `https://arxiv.org/abs/2112.04359`.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

World Wide Web Consortium (W3C). RDF 1.2 Primer, 2024. URL `https://w3c.github.io/rdf-primer/spec/`. Accessed: 2024.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis Ioannidis, Karthik Subbian, James Y Zou, and Jure Leskovec. Stark: Benchmarking llm retrieval on textual and relational knowledge bases. *Advances in Neural Information Processing Systems*, 37:127129–127153, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*. OpenReview.net, 2019.

Xiaoran Xu, Wei Feng, Yunsheng Jiang, Xiaohui Xie, Zhiqing Sun, and Zhi-Hong Deng. Dynamically pruned message passing networks for large-scale knowledge graph reasoning. In *ICLR*. OpenReview.net, 2020a.

Xiaoran Xu, Wei Feng, Yunsheng Jiang, Xiaohui Xie, Zhiqing Sun, and Zhi-Hong Deng. Dynamically pruned message passing networks for large-scale knowledge graph reasoning. In *International Conference on Learning Representations*, 2020b. URL `https://openreview.net/forum?id=rkeuAhVKvB`.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*, 2015a.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*, 2015b.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 40–48. JMLR.org, 2016.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9236–9260, 2024.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. L2-gcn: Layer-wise and learned efficient training of graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2127–2135, 2020.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. GraphSAINT: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJe8pkHFwS.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhao Zhang, Fuzhen Zhuang, Hengshu Zhu, Zhi-Ping Shi, Hui Xiong, and Qing He. Relational graph neural network with hierarchical attention for knowledge graph completion. In *AAAI*, pages 9612–9619. AAAI Press, 2020.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38, 2024a.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*, 2024b.

Hattie Zhou, Ankit Vani, Hugo Larochelle, and Aaron Courville. Fortuitous forgetting in connectionist networks. In *International Conference on Learning Representations*, 2022.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal A. C. Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. *CoRR*, abs/2106.06935, 2021. URL `https://arxiv.org/abs/2106.06935`.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.

Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, and Quanquan Gu. Few-shot representation learning for out-of-vocabulary words. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.