# Contents

# Chapter 5

# Improving Language Plasticity via Pretraining with Active Forgetting

*A version of this work was previously presented at a peer-reviewed conference. Please refer to [Chen et al., 2023] for full citation.*

Reality is full of constantly changing details. To navigate such dynamism, intelligent agents must adapt to new information in *real time*. This requires mechanisms that support flexible knowledge integration. *Active forgetting* (Chapter 4) appears to be one such mechanism: by actively forgetting historical node states resulted from previous message-passing computation, factorization-based models – representatives of the structured paradigm – can learn to accommodate new entity nodes in knowledge graphs, weaving them into the fabric of existing knowledge. At its core, active forgetting manifests an emergent principle of *destructuring*:

> To remain adaptable in changing environments, intelligent units (e.g., agents, models, humans) must not only construct knowledge, but also deliberately dismantle parts of it.

Increasingly, similar manifestations of such intentional destructuring have been identified across domains including but not limited to psychology, neuroscience, education, and artificial intelligence [Levy et al., 2007, Barrett and Zollman, 2009, Hardt et al., 2010, 2013, Anderson and Hulbert, 2021, Nikishin et al., 2022, Zhou et al., 2022, Ramkumar et al., 2023], reinforcing the idea that intelligence, especially its fluid side [Cattell, 1963, Horn and Cattell, 1966, Brown, 2016, Kent, 2017], relies as much on destructuring as on structuring. Structuring provides the foundations for consistent reasoning and repeatable knowledge serving. Destructuring, on the other hand, overcomes outdated and overly-rigid structures.

One of the key challenges in materializing the destructuring principle is to find the targets to dismantle. For natural intelligence, the targets of destructuring can be both cognitive and psychic structures. For instance, dismantling entrenched associative thinking patterns can lead to novelty in idea generation [Horan, 2009], while breaking down rigid psychic structures increases mental mobility, turning behavioural rigidity into feeling, thinking, and action [Sandell, 2019]. Similarly, inhibition of linguistic structures from one's native language plays an important role in acquiring a second language [Levy et al., 2007, MacWhinney, 2005, Schmid, 2017].

For artificial intelligence, the targets of destructuring remain understudied. Partially because scaling model sizes is the focus right now as it is more prominent in improving benchmark numbers. However, as more and more inappropriate behaviours by these models are exposed [Farquhar et al., 2024, Shumailov et al., 2024], it becomes more and more important to underpin these inappropriate structures inside the models. Chapter 2 and 3 show that certain structures are stored in the embeddings and their interactions with other layers in both the structured and unstructured learning paradigms. This perspective offers tangible structural underpinnings to the embedding layer. Chapter 4 further explains the role of embedding and chose them as the targets for destructuring, with evidences showing this helps models accommodate new entities in the knowledge graphs. While the findings from Chapter 4 are limited to the structured learning paradigm, an important question arises: can similar destructuring techniques benefit models operating in the unstructured paradigm. Specifically, we ask *can pretrained language models, the predominant tools for constructing knowledge engines from unstructured data sources, benefit from destructuring techniques?*

## 5.1 Towards Language Model Plasticity

Pretrained language models (PLMs) have been swiftly reshaping the landscape of natural language processing (NLP) by improving upon standardized benchmarks across the board [Radford and Narasimhan, 2018, Devlin et al., 2019, Liu et al., 2019b, Brown et al., 2020]. They are often regarded as the Swiss Army knife of the unstructured paradigm for building general knowledge engines. At their core, they acquire knowledge by ingesting large datasets and store this knowledge in their parameters during pretraining. Using finetuning or prompting [Brown et al., 2020], such knowledge can then be applied to downstream applications, such as semantic analysis, question answering, writing assistance, coding companion, and many others.

Despite their success, PLMs still have a number of shortcomings [Weidinger et al., 2021, 2022]. In particular, it requires massive data and computation to pretrain them [Gururangan et al., 2020, Kaplan et al., 2020, Hernandez et al., 2021, Hu et al., 2021, Touvron et al., 2023b]. Naively retraining a new PLM to accommodate every lingual space shift[1] would be prohibitively expensive. This makes it a highly relevant research target to create PLMs that can be efficiently adapted to new lingual spaces.

While forgetting in the context of both human and machine learning is often perceived as something negative (for instance in the case of catastrophic forgetting where learning new tasks overwrites the old knowledge [McCloskey and Cohen, 1989, Ratcliff, 1990, Kirkpatrick et al., 2017]), recent works have shown that for artificial neural networks, forgetting can also play a *positive* role in increasing their "plasticity", such as improving generalization to unseen data [Zhou et al., 2022, Chen et al., 2022, Igl et al., 2021], enabling learning in low-data regimes [Alabdulmohsin et al., 2021, Taha et al., 2021], or counteracting primacy bias [Nikishin et al., 2022, D'Oro et al., 2023]. Although these pioneering works in continual learning do not explicitly define model plasticity, they in essence share a common goal across different tasks and models: improving a model's ability to remain stable while adapting flexibly to drastically changing inputs, addressing the *stability-plasticity dilemma*. Given these developments and their insights, in this work, we explore if we can draw upon forgetting techniques as a mechanism to improve *pretraining* and imbue PLMs with similar benefits in model plasticity.

It is well established in the NLP community that models struggle to generalize across languages without substantial intervention [Conneau et al., 2020, Pfeiffer et al., 2020, 2022, Ansell et al., 2022], which is especially true for low-resources languages. We thus

[1]We use the term *lingual space shift* to describe changes in language usage between pretraining and the target downstream application, caused by factors such as language change, time evolution, or domain variation. A model with high *language plasticity* would quickly adapt to these shifts.

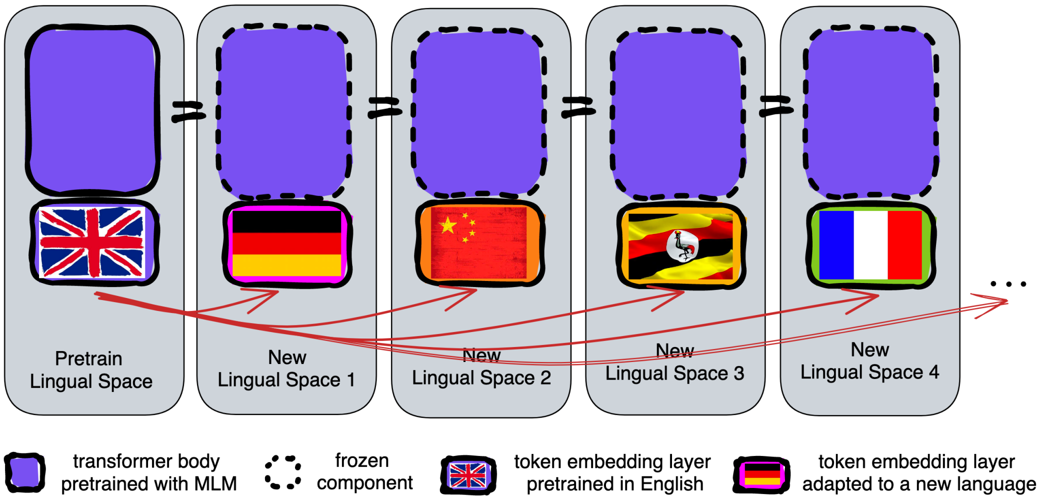


Figure 5.1: *Rewiring* via relearning token embeddings: where the transformer body (the purple part) is "frozen" and reused for a new language, but the token embeddings are relearned to suit the new language.

see this as a promising testing ground for forgetting techniques. Our focus is on the input layer of the PLM, the *token embedding layer*, as learning it has been shown to be highly effective when adapting between languages [Artetxe et al., 2020].

Concretely, we propose to introduce *active forgetting* mechanism into the pretraining phase, which resets token embeddings at regular intervals, while leaving all other parameters untouched throughout pretraining. We study whether this forgetting approach creates a PLM that can easily *rewire* (Figure 5.1) to an unseen (possibly distant) language. Intuitively, resetting embeddings forces the transformer body to re-derive reasoning each time instead of relying on memorized shortcuts. Through repetition, the body learns more abstract, high-level reasoning. A model with greater abstraction can easily transfer across languages, since high-level reasoning is more language-agnostic.

Our zero-shot evaluations on several cross-lingual transfer benchmarks show that for cases where unlabeled adaptation corpus for the unseen language has as few as 5 million tokens (a low-data regime), forgetting PLMs outperforms the baseline by large margins: average gains of $+21.2\%$ on XNLI, $+33.8\%$ on MLQA, and $+60.9\%$ on XQuAD. In addition, models pretrained using active forgetting converge faster during language adaptation. Finally, we find that active forgetting is especially beneficial for languages that *are* *distant from* English, such as Arabic, Hindi, Thai, and Turkish. Implementation-wise, the method does not introduce significant overhead to the already complex pretraining process, making it a cost-efficient way to promote a meta-learning-like effect. For those interested in details, the code is available at `https://github.com/facebookresearch/language-model-plasticity`.

## 5.2 Literature Review: Forgetting, its Positive Roles, and Cross-lingual Transfer

### 5.2.1 Forgetting and its Positive Role

The common perception of forgetting is that it implies weak memory and a loss of acquired knowledge, thus it is often regarded as a sign of *un-intelligence* or an undesirable property. In neural networks, *catastrophic forgetting* [McCloskey and Cohen, 1989, Ratcliff, 1990, Kirkpatrick et al., 2017] is portrayed as a forgetting phenomenon where neural networks lose the ability to predict old patterns after new inputs alter their weights. Forgetting, in this context, has negative consequences, as the new knowledge overwrites the prior valuable knowledge. Plenty of prior research strives to overcome catastrophic forgetting and enable continual learning [Schmidhuber, 2013, Kirkpatrick et al., 2017, Lopez-Paz and Ranzato, 2017, Shin et al., 2017, Schwarz et al., 2018, Mallya and Lazebnik, 2018, Parisi et al., 2019, Rolnick et al., 2019, Beaulieu et al., 2020, Veniat et al., 2020, Gaya et al., 2023, Khetarpal et al., 2022].

Our work differs from the above ones in that our subject is *intentional forgetting* rather than passive forgetting and its associated negative impact. To put it in another way, we seek to understand how forgetting – if purposely incorporated as an active process during training – might *help* new learning. Similar positive roles of forgetting have been discussed in the literature. Specifically, Pastötter et al. [2008] demonstrate forgetting enhances the learning of new information by resetting the encoding process and holding the attention at high levels; Levy et al. [2007] show that it helps second language acquisition by inhibiting the native language; Barrett and Zollman [2009] find it promote the emergence of an optimal language by preventing partial success from reinforce suboptimal practice. Nørby [2015] further suggests forgetting serves adaptive functions, helping people regulate emotions, acquiring knowledge and staying attuned to the context. More recently Anderson and Hulbert [2021] reviews evidence on active forgetting by prefrontal control and shows how it can adapt the memory to suit either emotional or cognitive goals.

### 5.2.2 Forgetting via Partial Neural Weights Reset

In neural networks, forgetting can be instantiated in many forms. A simple way is to reset subsets of parameters before the next round of learning. Iterations of such resetting have been shown to benefit generalization with low compute and low data for computer vision tasks [Frankle and Carbin, 2019, Alabdulmohsin et al., 2021, Taha et al., 2021, Ramkumar et al., 2023]. More recently, Zhou et al. [2022] demonstrate a similar forgetting strategy helps image classification and language emergence. Closely linked to the method in this chapter, Chapter 4 forget node embeddings in order to truncate infinite message-passing among nodes and thereby aid new graph reasoning with new nodes. Our work uses similar forgetting mechanism over token embeddings, improving new language reasoning with new tokens. As far as we know, *we are the first to bring forgetting into pretraining and demonstrate that forgetting pretraining boosts linguistic plasticity*. A relevant thread in reinforcement learning (RL) research studies the plasticity loss phenomenon [Lyle et al., 2023, Nikishin et al., 2023]. Recent work explores similar forgetting approaches to improve plasticity. Igl et al. [2021] periodically reset the current policy by distilling it into a reinitialised network throughout training. Intuitively, this releases network capacity storing suboptimal policies and opens up space for the yet-to-be-discovered optimal (final) policy. Simpler methods just reset an agent's last layers [Nikishin et al., 2022], preventing overfitting to early experiences and *primacy bias*. Resetting parameters also improves sample efficiency by allowing more updates per environment interaction [D'Oro et al., 2023].

### 5.2.3 Cross-lingual Transfer for Pretrained Language Models

Pretraining on multilingual data makes PLMs multilingual [Conneau et al., 2020] but has downsides like needing large multilingual corpus with appropriate mixing, potential interference among languages, and difficulty of covering all languages. Alternatively, the line of research on cross-lingual transfer makes PLMs multilingual by extending

English-only PLMs to other languages. Artetxe et al. [2020] demonstrate possibility of such extension by relearning the embedding layer with unsupervised data from the new language. Marchisio et al. [2023] further increase computational efficiency using a mini-model proxy. Liu et al. [2023a] use a similar partial reset-reinit approach in vision-language settings. Approaches based on adapters and sparse finetuning have also been proposed [Pfeiffer et al., 2020, 2022, 2021, Ansell et al., 2022]. Adapters are bottleneck layers (usually placed after the feedforward layers) that add extra capacity when adapting to a different task or language. Our proposed forgetting mechanism can be applied to adapter-based methods as we can allow forgetting to happen in the adapter layers. The current choice of forgetting embeddings keeps the architecture intact and incurs no additional hyperparameter tuning, allowing us to understand the fundamental capability of forgetting pretraining.

## 5.3 Rewiring PLMs for New Languages

Using unlabeled data, Artetxe et al. [2020] demonstrates possibility of rewiring a monolingual PLM to a new language; they propose to relearn the embedding layer for the new language while keeping all the other parameters frozen. The underlying assumption is that the token embedding layer and the transformer body (the non-token-embedding parameters) divide up the responsibility in a way that the former handles language-specific lexical meanings, while the latter deals with high-level general reasoning. Hence, rewiring an English PLM for a new language boils down to separately adapting the former with unlabelled data in the new language and the latter with English task data. The procedure can be summarized as follows:

1. Pretrain: A transformer-based model is pretrained on an *English* corpus. In our experiments, we choose to pretrain RoBERTa-base Liu et al. [2019b], a 12-layer transformer-based model, on English CC100 [Conneau et al., 2020].

2. Language Adapt: The token embedding layer is finetuned using unlabelled data in the new language, while the transformer body is frozen.

3. Task Adapt: The transformer body is finetuned using downstream task data in English, while the token embedding layer is frozen.

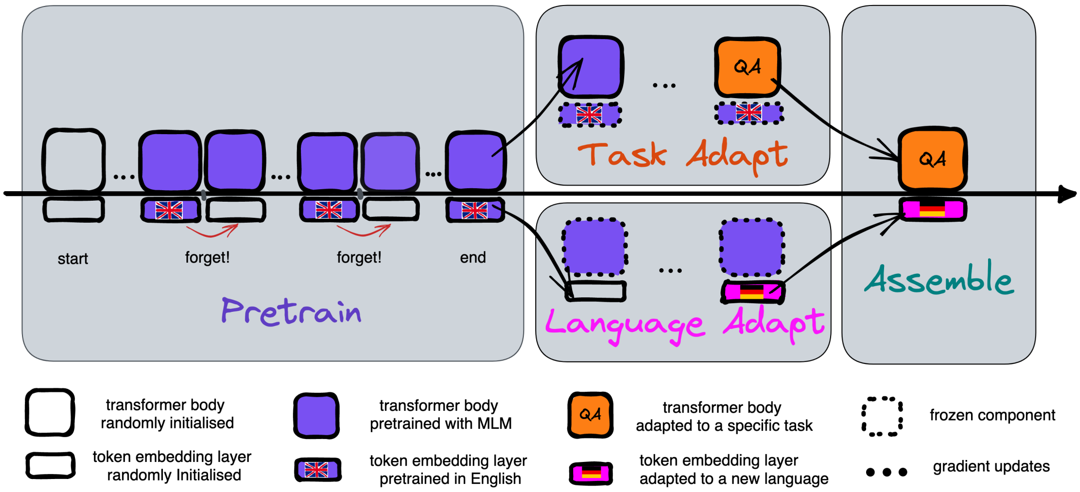


Figure 5.2: Unsupervised zero-shot cross-lingual transfer. **Left**: in the pretrain stage, we compare standard pretraining with forgetting pretraining, where the token embeddings are actively forgotten at a regular interval while the transformer body is learned as the standard pretraining. **Middle**: the task adapt stage and the language adapt stage separately adapt the transformer body using English task data and the token embeddings using unlabelled data in the new language. **Right**: the assemble stage reassemble the adapted body and token embedding layer into a usable PLM.

4. Assemble: The final model is assembled by taking the adapted token embedding layer from stage 2 and the adapted transformer body from stage 3.

## On The Difficulty of Rewiring PLMs via Relearning Token Embeddings

While the above procedure [Artetxe et al., 2020] offers a general framework for rewiring a monolingual PLM with unlabelled data in the new language, it is unclear how efficient such rewiring can be, including both sample efficiency and computational efficiency. To better understand the difficulty of rewiring PLMs via relearning the token embeddings, we design an experiment where we relearn the token embedding layer using varying amounts of adaptation data. For illustration purpose, we pick English as the pseudo "adaptation language" as its dataset is large enough to bootstrap a series of sub-datasets with varying quantity.

We create subsets with $[1\text{K}, 10\text{K}, 100\text{K}, 1\text{M}, 5\text{M}, 10\text{M}, 100\text{M}, 1\text{B}, 10\text{B}]$ tokens and

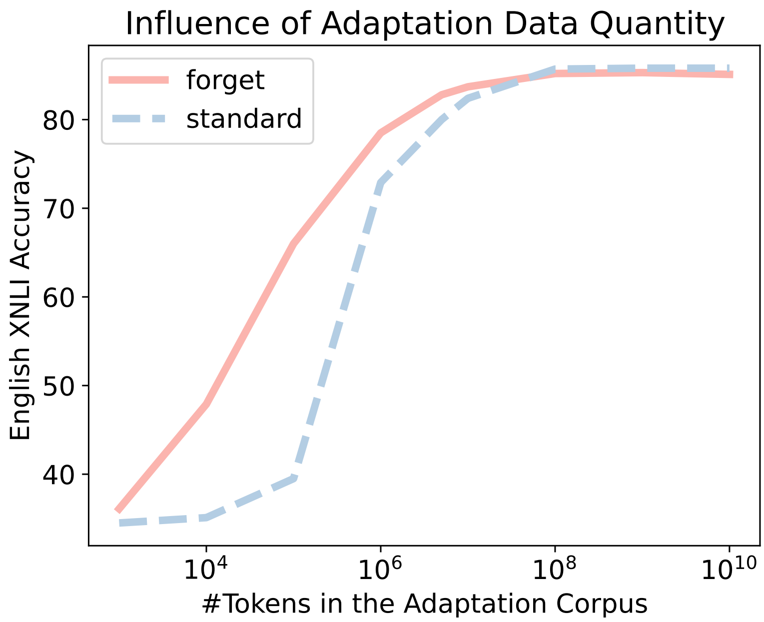


Figure 5.3: The rewiring performance for standard PLMs (blue dashed line) drops drastically if the adaptation tokens $\leq$ 10M.

relearn the English embeddings while keeping the transformer body frozen.

The dashed blue line in Figure 5.3 summarizes the influence of the adaptation data quantity on the quality of the rewired PLMs (relearned embeddings assembled with the English NLI task body). We can see that the standard PLMs are easy to rewire if there is enough adaptation data. However, if the adaptation corpus contains fewer than 10 million tokens, the performance of the rewired standard PLMs (the blue dashed line in the figure) drops drastically as the adaptation data quantity goes down, from near 80 to around 35, a random-guessing level for NLI tasks. This motivates us to create more rewirable PLMs, i.e. PLMs with more plasticity so that the rewiring process can be faster and consume less data.

## 5.4 Pretraining with Active Forgetting

Recent works have shown that incorporating forgetting through iterative weights resetting can increase the "plasticity" of neural networks, enabling them to learn from small data and generalize better to unseen data in supervised learning [Alabdulmohsin et al., 2021, Taha et al., 2021, Zhou et al., 2022]. Building on these efforts, we study if we can bring such forgetting into the pretrain stage so that the resulting PLM would have more plasticity, allowing easier rewiring to new languages.

**Our Hypothesis.** In effect, when Artetxe et al. [2020] relearned the token embedding layer, the reinitialisation of the embeddings can be seen as forgetting applied *once* at the start of the language adapt stage. However, the PLM (specifically the transformer body) has never encountered forgetting before this stage and may struggle to handle this new situation. Without early exposure to forgetting, the PLM might suffer from slow recovery caused by forgetting before eventually benefiting from it. This inefficiency also implies a lack of plasticity in the Transformer architecture. During standard pretraining, token embeddings in the Transformer can encode excessive structures tied to the specifics of their training languages so that other parts of these models become overly rigid to the linguistic characteristics of the training language. The learning of a new lexical embedding layer in a PLM henceforth consumes lots of data in new languages along with long training horizons as shown in Section 5.3. In this chapter, to ensure swift learning of the new languages with both high sample efficiency and convergence rate, we argue that the PLM must be exposed to forgetting during pretraining, allowing itself to maximize the positive impact of forgetting and minimizing the cost of recovery.

**Our Method.** With this hypothesis in mind, we propose to add an *active forgetting* mechanism to the pretraining procedure, which resets the token embedding layer periodically as described in Algorithm 5. Concretely, the forgetting mechanism operates by intentionally clearing the weights of the embedding layer, which stores the static representations for all tokens, and reinitialising them to a new set of random values every $K$ gradient updates. Since pretraining involves advanced training strategies, like optimizers with states and learning rate schedulers, we also reset them together with the token embedding layer. We refer to language models pretrained with such active forgetting mechanism as *forgetting PLMs*, in contrast to *standard PLMs* which are pretrained in a standard way. The pretraining loss curve of a forgetting PLM is episodic (Figure 5.4), like in reinforcement learning or meta-learning. This episodic learning demonstrates that the active forgetting mechanism can introduce diversity without requiring

**Algorithm 1:** Active Forgetting Mechanism. The learning of token embedding layer is reset every $K$ updates.

**Input:** $K$: interval between two consecutive forgetting;
$n_{\text{body/emb}}$: current effective number of updates for the body or the token embedding layer;
$\alpha_{\text{body/emb}}$: current learning rate for the body or the token embedding layer;
$P^n_{\text{body/emb}}$: parameters after the $n^{\text{th}}$ update for the body or the token embedding layer;
$O^n_{\text{body/emb}}$: optimizer states after the $n^{\text{th}}$ update for the body or the token embedding layer;
$\Theta$: randomly initialised embedding parameters, each element drawn from $\mathcal{N}(0, 0.02)$;
$f$: function that computes the gradients w.r.t. the parameters using the sampled data;
$g$: function that updates the parameters based on the gradients (e.g., one step in Adam optimizer);
$s$: function that updates the learning rate (e.g., one step in the polynomial learning rate scheduler).

**Output:** The updated parameters and optimizer states:
$P^{(n)} = \{P^{(n)}_{\text{emb}}, P^{(n)}_{\text{body}}\}$,
$O^{(n)} = \{O^{(n)}_{\text{emb}}, O^{(n)}_{\text{body}}\}$.

$n_{\text{emb}} \leftarrow n \mod K$;
$\alpha_{\text{body}} \leftarrow s(n_{\text{body}})$ `// Adjust learning rate for body based on` $n$;
$\alpha_{\text{emb}} \leftarrow s(n_{\text{emb}})$;
$G^{(n)} \leftarrow f(P^{(n-1)}, \cdot)$ `// Compute all gradients;`
$P^{(n)}_{\text{body}}, o^{(n)}_{\text{body}} \leftarrow g(G^{(n)}_{\text{body}}, P^{(n-1)}_{\text{body}}, o^{(n-1)}_{\text{body}}, \alpha_{\text{body}}, n)$ `// Update the transformer body;`
**if** $n_{emb} == 0$ **then**
  $P^{(n)}_{\text{emb}} \leftarrow \Theta$ `// Reset token embeddings and relevant optimizer states;`
  $o^{(n-1)}_{\text{emb}} \leftarrow 0$;
$P^{(n)}_{\text{emb}}, o^{(n)}_{\text{emb}} \leftarrow g(G^{(n)}_{\text{emb}}, P^{(n-1)}_{\text{emb}}, o^{(n-1)}_{\text{emb}}, \alpha_{\text{emb}}, n_{\text{emb}})$ `// Update the token embeddings;`

actual new data. Each forgetting event kind of "branches out" a novel environment for the model to explore, as if initiating a new episode of learning.

**Research Questions.** To further examine the proposed forgetting mechanism, we compare *forgetting PLMs* and *standard PLMs* on sample efficiency and convergence speed during language adapt, two key aspects of model plasticity. Our research investigates:

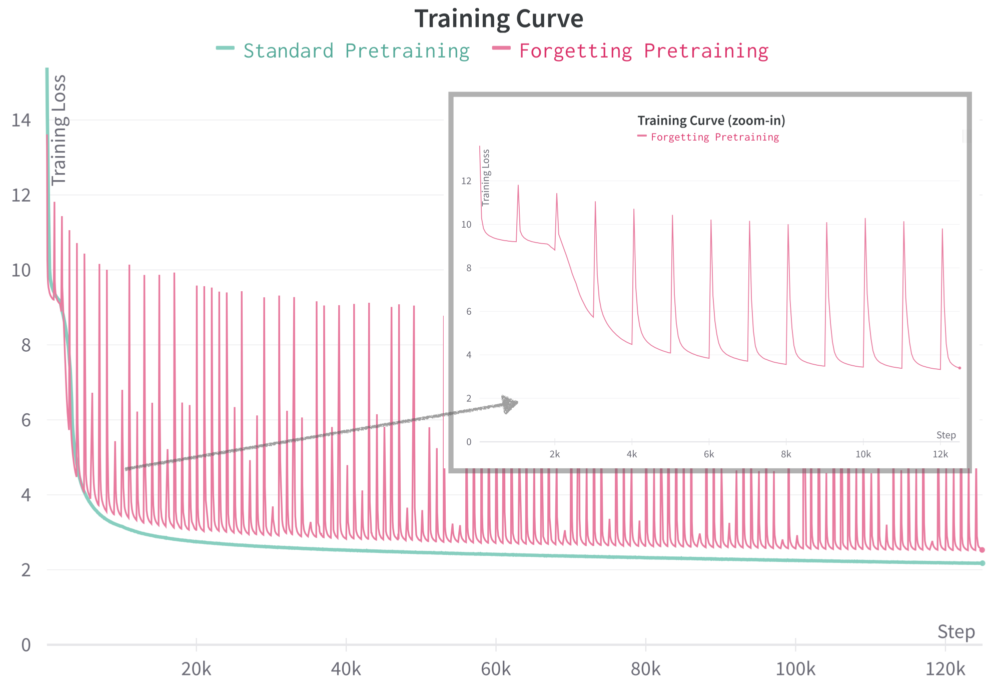


Figure 5.4: Pretraining losses of forgetting and standard language models. The forgetting mechanism brings an episodic pattern into the loss curve: every embedding forgetting produces a loss spike, from which the model learn to recover. Through such repeats of forget-relearn, the model gets used to learn new embeddings from scratch.

- RQ1: Real-world low-resource languages often have scarce data for adapting models. Does pretraining with active forgetting impart enough plasticity to forgetting PLMs, enabling them to learn new languages even with such limited data?
- RQ2: Deploying PLMs frequently encounters computational limitations. Endowed with more plasticity, can forgetting PLMs reduce adaptation time for such low-compute scenarios?
- RQ3: New languages may be very similar or different from pretraining languages. Does this similarity/difference impact the relative benefit of forgetting PLMs over standard PLMs?

## 5.5 Experiments

To evaluate the effectiveness of forgetting PLMs and address RQ1-RQ3, we conduct experiments on several cross-lingual transfer benchmarks.

### 5.5.1 Experimental Setup

In our work, we closely follow the setup in Artetxe et al. [2020] and Marchisio et al. [2023]. Our pretraining model is RoBERTa-base, a standard 12-layer transformer-based language model. We trained for each language a sentencepiece tokenizer [Kudo and Richardson, 2018] with a vocabulary size of 50K over the corresponding data subsets in CC100. The model was pretrained with the English subset of the CC-100 dataset. The pretraining process consists of 125K updates, with a batch size of 2048. We used a learning rate scheduler with linear decay and an initial learning rate of $7\mathrm{e}-4$, with 10K warm-up updates. Checkpoints were saved every 500 updates. Since longer pretraining consistently led to better validation perplexities in our experiments, we chose the final pretraining checkpoint (step 125K) whenever possible for optimal performance. Since the final checkpoint might coincide token embeddings reset in forgetting pretraining, we instead chose the closest checkpoint that has the best validation perplexity. This ensured that we selected the best pretrained checkpoints for both approaches based on when they achieved their optimal validation perplexities. We set the frequency of forgetting $\mathrm{K} = 1000$ and used a clip-norm of 0.5.

During the language adapt stage, we kept most of the hyperparameters the same as for pretraining. We finetuned the token embedding layer while keeping the others frozen, as described in Section 5.3. This differs from the pretraining setup, where all parameters are learnable to maximize learning speed. In contrast, the finetuning setup is intended to mimic how humans might typically relearn word meanings: by updating embeddings while keeping the rest of the system fixed. Note that *no* forgetting happens during this stage because we want the models to learn the new languages as well as possible. In the task adapt stage, both models were finetuned for 10 epochs on the English task data, specifically MultiNLI [Williams et al., 2018] for the NLI task and SQUAD Rajpurkar et al. [2016] for the QA task. After the assemble stage, we evaluate the zero-shot performance of the assembled model on XNLI [Conneau et al., 2018], a cross-

Table 5.1: Accuracy comparison of forgetting and standard PLMs on the XNLI dataset (table continues).

| **Method** | **vi** | **sw** | **es** | **bg** | **de** | **fr** | **el** | **ru** |
|---|---|---|---|---|---|---|---|---|
| Standard | **65.8** | 55.6 | 68.0 | 65.5 | 62.2 | 63.5 | 63.1 | 56.9 |
| Forgetting | 62.8 | **59.5** | **74.0** | **71.7** | **68.5** | **71.2** | **70.8** | **65.8** |
| Gain(%) | $-4.6$ | $+7.0$ | $+8.8$ | $+9.5$ | $+10.1$ | $+12.1$ | $+12.2$ | $+15.6$ |

lingual NLI task, along with XQuAD [Artetxe et al., 2020] and MLQA [Lewis et al., 2020a], two cross-lingual QA tasks. We report the NLI accuracy and QA F1 on the test sets.

Our experiments were implemented using fairseq [Ott et al., 2019]. The pretraining and language adaptation experiments were conducted on 32 Tesla V100 GPUs (each with 32 GB memory) and took approximately 24-36 hours to complete. The time taken for both stages were quite close to each other even though the latter only involved tuning the embeddings. This demonstrates the importance of reducing the computational cost of the language adaptation stage.

Differing from prior work [Artetxe et al., 2020, Marchisio et al., 2023], we focus on language adapt in low-data regimes. We simulate low-resources scenarios by limiting the adaptation data for each downstream language to only 5M subword tokens from CC100. This is in contrast with conventional setups, where all the tokens in the corresponding languages in CC100 are used for language adaptation. As Table C.2 shows, such setups consume several orders of magnitude more data than our 5M-token setup; for instance, the Swahili CC100 subset contains 345M tokens, roughly 69 times larger than our corpus, and the Russian subset contains 34.9B tokens, roughly 6,980 times larger. Therefore, PLMs that can successfully learn new languages with rich data under traditional setups may struggle to do so with our limited 5M-token corpus.

### 5.5.2 RQ1: Forgetting PLMs Work Better in Low-Data Regimes

Standard PLMs struggle in low-data language adaptation, dropping from 86.1 English NLI accuracy to just 53.3 average accuracy on XNLI with limited 5M token adaptation data. Compared to prior work which uses full data from Wikipedia [Artetxe et al., 2020]

Table 5.2: Accuracy comparison of forgetting and standard PLMs on the XNLI dataset (table continued). On average, forgetting achieve a 21.2% relative gain in accuracy compared to standard across the languages tested, where averaged relative gain = $\frac{\sum_{x \in \{\text{languages}\}} \text{Relative Gain of } x}{\text{\#Languages}}$.

| **Method** | **zh** | **ur** | **hi** | **tr** | **ar** | **th** | **Avg** | **en** |
|---|---|---|---|---|---|---|---|---|
| Standard | 53.2 | 36.8 | 39.7 | 38.9 | 41.2 | 35.3 | 53.3 | **86.1** |
| Forgetting | **63.5** | **45.8** | **52.9** | **52.7** | **59.5** | **59.7** | **62.7** | 85.1 |
| Gain(%) | +19.4 | +24.5 | +33.2 | +35.5 | +44.4 | +69.1 | +21.2 | −1.2 |

Table 5.3: F1-score comparison of forgetting and standard PLMs on MLQA. On average, forgetting PLMs achieve a 33.8% relative gain in F1 compared to standard PLMs across the languages tested, where averaged relative gain = $\frac{\sum_{x \in \{\text{languages}\}} \text{Relative Gain of } x}{\text{\#Languages}}$.

| **Method** | **es** | **vi** | **de** | **zh** | **hi** | **ar** | **Avg** | **en** |
|---|---|---|---|---|---|---|---|---|
| Standard | 49.4 | 38.3 | 45.3 | 34.1 | 17.7 | 20.8 | 34.3 | **78.9** |
| Forgetting | **55.3** | **45.0** | **53.4** | **43.0** | **28.8** | **34.7** | **43.4** | 78.3 |
| Gain(%) | +12.0 | +17.6 | +17.8 | +26.2 | +62.5 | +67.0 | +33.8 | −0.8 |

or from CC100 [Marchisio et al., 2023], the average accuracy on XNLI drops about 18% (from 66.8/66.3 to 53.3). This indicates standard PLMs are not coping well with the low-data regime. In contrast, forgetting PLMs achieve decent 62.7 average XNLI accuracy, a +21.2% relative gain over standard PLMs, as shown in Table 5.2.

Forgetting PLMs also outperform standard PLMs on MLQA and XQuAD, with average F1 relative gains of +33.8% and +60.9% across languages, as respectively demonstrated in Table 5.3, Table 5.4 and Table 5.5. Across NLI and QA tasks, forgetting PLMs consistently surpass standard PLMs in low-data regimes. Why do forgetting PLMs handle the low-data regime better? We hypothesize this is because forgetting PLMs are more robust to different embedding initialisations. They encode more universal knowledge in the transformer body. Standard PLMs may encode more "shortcut" knowledge relying on certain embedding initialisations. In low data, standard PLMs cannot adjust embeddings towards shortcut routes without access to enough data. Forgetting PLMs do not rely on shortcuts so perform better.

Table 5.4: F1-score comparison of forgetting and standard PLMs on XQuAD (table continues). On average, forgetting PLMs achieve a 60.9% relative gain in F1 compared to standard PLMs across the languages tested, where averaged relative gain $= \frac{\sum_{x \in \{\text{languages}\}} \text{Relative Gain of } x}{\text{\#Languages}}$.

| **Method** | **vi** | **es** | **ru** | **de** | **el** | **zh** |
|---|---|---|---|---|---|---|
| Standard | 49.7 | 57.7 | 49.4 | 50.9 | 48.5 | 32.4 |
| Forgetting | **52.9** | **64.6** | **56.5** | **60.9** | **59.9** | **43.7** |
| Gain(%) | +6.4 | +12.0 | +14.5 | +19.7 | +23.6 | +34.6 |

Table 5.5: F1-score comparison of forgetting and standard PLMs on XQuAD (table continued). On average, forgetting PLMs achieve a 60.9% relative gain in F1 compared to standard PLMs across the languages tested, where averaged relative gain $= \frac{\sum_{x \in \{\text{languages}\}} \text{Relative Gain of } x}{\text{\#Languages}}$.

| **Method** | **hi** | **ar** | **th** | **tr** | **Avg** |
|---|---|---|---|---|---|
| Standard | 21.4 | 22.2 | 15.4 | 13.0 | 36.1 |
| Forgetting | **33.3** | **38.7** | **38.4** | **41.4** | **49.0** |
| Gain(%) | +55.8 | +74.2 | +149.7 | +218.8 | +60.9 |

### 5.5.3 RQ2: Rewiring Forgetting PLMs Requires Fewer Updates

We are also interested in how quickly forgetting PLMs and standard PLMs can learn new languages. Figure 5.5 summarizes adaptation curves on XNLI, MLQA and XQuAD, with each point representing the averaged performance across all languages. In just 5K steps (4% of full adaptation), forgetting PLMs reach 57.8 accuracy on XNLI while standard PLMs struggle at random guessing levels of 37.2. Similar trends hold for MLQA and XQuAD. After 5K steps, forgetting PLMs achieve 92% of their full performance on XQuAD versus just 53% for standard PLMs (see the last plot in Figure 5.5).

Why do forgetting PLMs converge faster? We hypothesize it is because periodical embedding resetting forces the body to gradually locate itself on a particular manifold, where it can easily cooperate with new embeddings. This makes the body encourage larger embedding updates when adapting to new languages. Active forgetting simulates

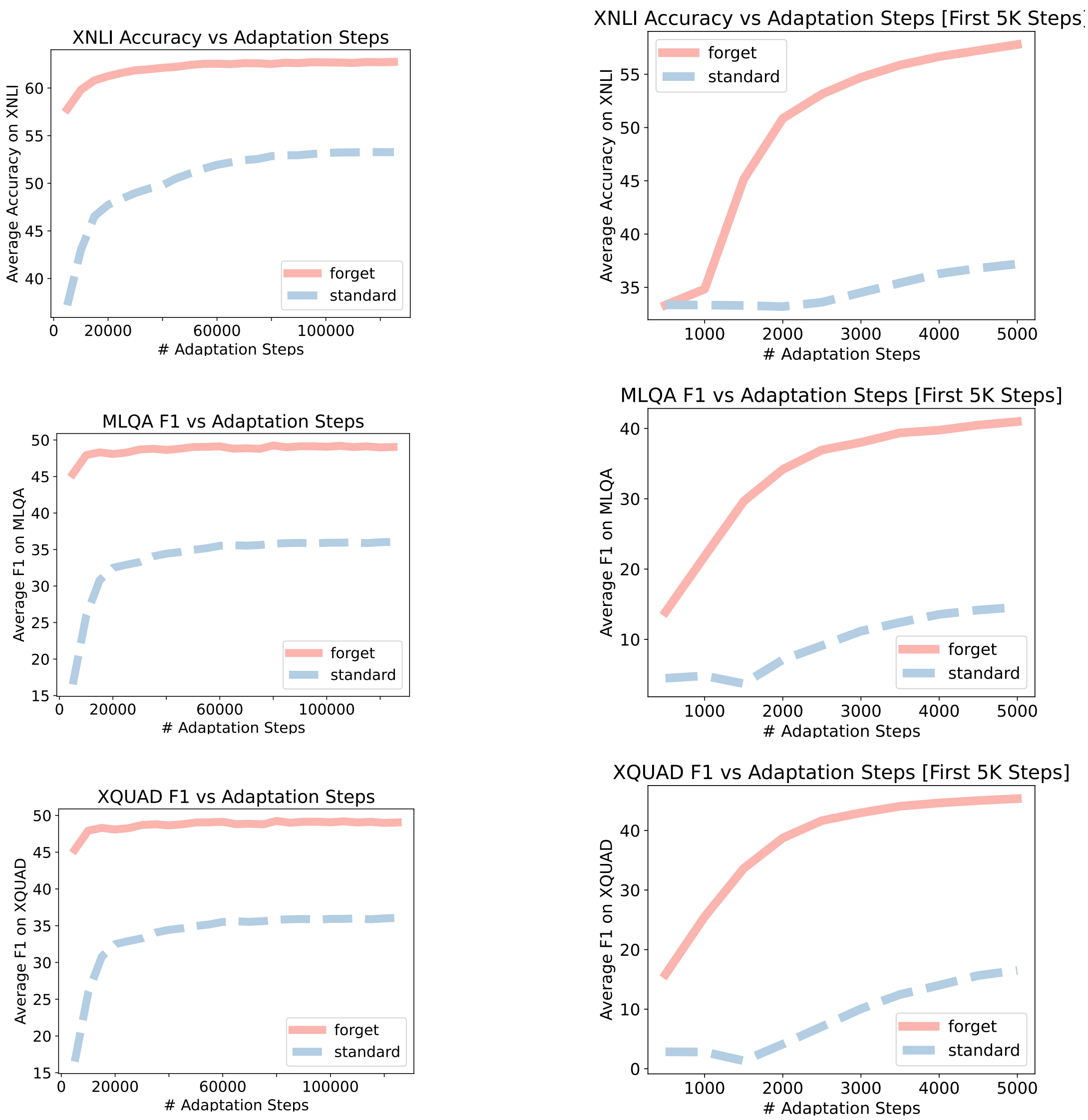


Figure 5.5: Adaptation curves on XNLI, MLQA, and XQuAD. Numbers aggregated across languages. The first row contains the full adaptation curves, which comprises 125K adaptation steps. The second row contains the zoom-in versions of curves for the first 5K adaptation steps. Forgetting PLMs converge faster than standard PLMs; for instance, on XQuAD (the last plot), forgetting PLMs reach 92% of their final performance within 5K updates, while standard PLMs only reached 53% of their final performance at that point.

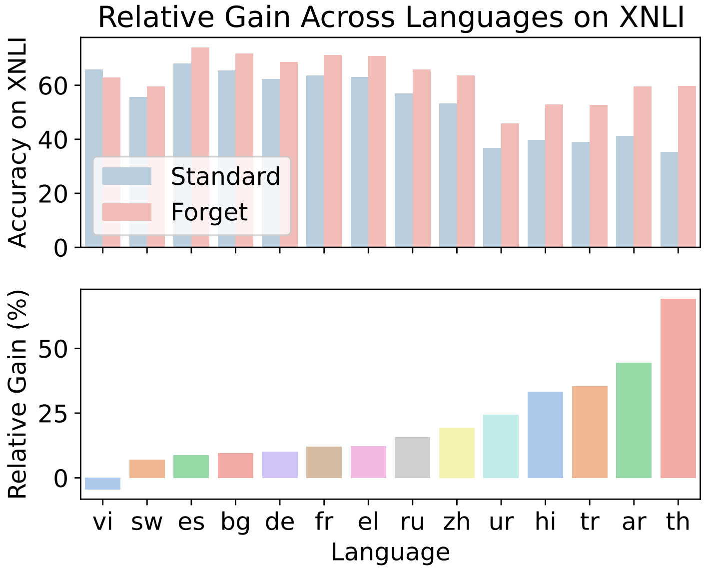


Figure 5.6: Relative gains of forgetting PLMs over standard PLMs across languages for XNLI. Forgetting yields substantial relative gains for languages like Arabic, Hindi, Thai, Turkish, and Urdu.

language switching during pretraining[2] introducing diversity without new data. This allows faster adaptation to real new languages.

### 5.5.4 RQ3: Distant Languages Benefit From Forgetting PLMs

We have primarily focused on discussing the averaged performance in the previous sections (Sec 5.5.2 and 5.5.3). In this section, we provide a more detailed comparison of language-specific performances between forgetting PLMs and standard PLMs on XNLI, MLQA, and XQuAD. To gain a deeper insight into which languages benefit the most from the use of forgetting, we present the relative performance changes across the languages in Figure 5.6 for XNLI and in Figure 5.7 for MLQA. For space reason, the results of XQuAD can be found in Figure C.1 in the appendix.

Across the spectrum of languages (Table C.1), we observe that forgetting provides greater benefits for languages distant to the pretraining language (English) in terms of language family, script and morphology. Specifically, forgetting brings large rela-

[2]Precisely, it simulates vocabulary swappings, causing drastic changes to the input of the body.

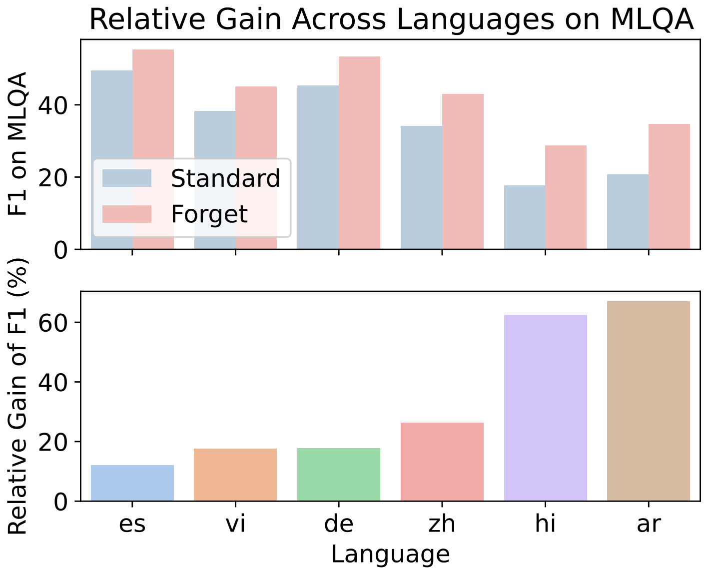


Figure 5.7: Relative gains of forgetting over standard across languages for MLQA. For languages closely related to English, such as German, the relative gains from forgetting are modest.

tive gains for languages such as *Arabic*, *Hindi*, *Thai*, *Turkish*, and *Urdu* compared to closer languages like *German*. Script seems important - forgetting helps Vietnamese and Swahili less despite their distance from English, likely due to the shared Latin script.

Languages that share a script with the pretraining language (e.g., English and German) tend to share subword tokens, enabling models to reuse learned embeddings and lexical patterns. This facilitates transfer and reduces the need to relearn low-level representations. In contrast, languages with different scripts (e.g., Arabic, Hindi, Thai) have minimal subword overlap and lack orthographic familiarity, making tokenization and representation learning more difficult. Script similarity, therefore, narrows the representational gap in cross-lingual transfer. Forgetting is more beneficial for script-divergent languages, as it enables the model to construct new, script-specific representations without interference from English.

Examining adaptation curves within the first 5K steps, forgetting PLMs reach substantially superior performance over standard PLMs for almost all languages except Urdu, while standard PLMs struggle at random guess levels (see Figure 5.8 and Section C.2). This demonstrates forgetting PLMs' ability to efficiently adapt to new languages,

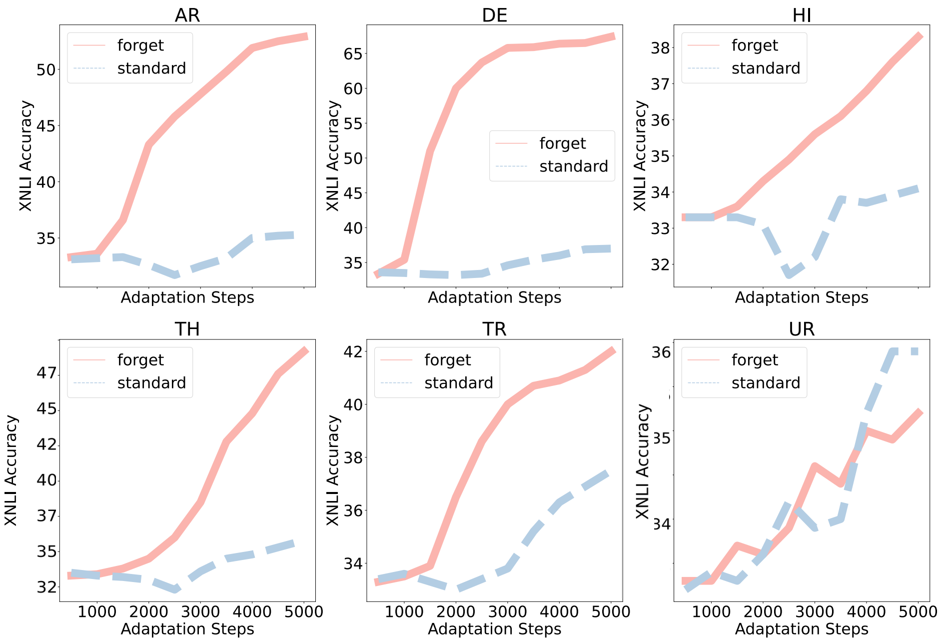


Figure 5.8: Adaptation curves on XNLI within 5K updates for individual languages: Bulgaria, Greek, Spanish, French, Russian, Swahili, Vietnamese and Chinese. For all languages except Urdu, the forgetting PLMs converge faster than the standard PLMs during the language adaptation stage.

particularly dissimilar ones, in low-data settings.

## 5.6 Discussion

**Summary** This chapter expands on the idea of *active forgetting*, a manifestation of the destructuring principle, and its potential impact on AI models. While Chapter 4 demonstrated the value of active forgetting in the structured paradigm for building general knowledge engines, this chapter applies it to unstructured paradigms, showing that *active forgetting* can improve pretrained language models by imbuing them with more linguistic plasticity. Experiments with RoBERTa show that models pretrained via active forgetting can better learn from **small data** while enjoying faster convergence during language adaptation, particularly for languages that are distant from English.

Most current efforts to build knowledge engines in the unstructured paradigm have been focusing on ingesting more data into larger models [Kaplan et al., 2020]. Accelerating techniques on both hardware and software sides are being developed to help us achieve such *structuring* of the reality (whether real or synthetic) into machine computation. On the other side, we as a community seem to have far fewer ideas on how we can **rewire** inappropriate structures from the models safely, timely, and relevantly [Weidinger et al., 2021, 2022, Kirk et al., 2024]. This chapter stands at the crossroad of structuring and destructuring, where we highlight the necessity of *destructuring* in its role for "machine plasticity" – a kind of freedom to delete built-in structures and rewire model behavior whenever needed. We **speculate** that destructuring may reduce the model's reliance on shortcut learning, where models depend on superficial cues rather than deeper structure [Geirhos et al., 2020]. By disrupting these shortcuts, destructuring could encourage the model to focus on more abstract patterns, potentially improving its ability to generalize to new environments.

The conclusion of this chapter, a dual focus on structuring and destructing, is surprising while providing a promising alternative to the scaling approach [Kaplan et al., 2020]. Destructuring can drive model evolution and rewire models to adapt to the dynamic world. Without this capacity for machine plasticity, we risk creating rigid AI systems that potentially trap their human users in outdated or biased "knowledge". A balance between structuring and destructuring opens the door to create more natural and flexible knowledge engines, ultimately supporting diverse AI applications that blend into our everyday life.

**Implications** Going beyond language adaptation, we argue that pretrained language models with more plasticity are a promising direction for future research, as they allow easier adaptation to various tasks, domains, languages and can evolve faster as the real world changes. Unlike symbolic methods, such as knowledge graphs, which can easily rewire a fact by modifying the corresponding knowledge triplet, current static PLMs are harder to rewire since changing one fact by updating model weights may disrupt multiple other facts without substantial post-hoc intervention. Improving the rewirability via forgetting pretraining thus can be seen as one way of imbuing PLMs with similar benefits as symbolic methods (making the resulted model more controllable i.e. can be modified with tiny cost), complementing the line of post-hoc model editing research [Mitchell et al., 2021, 2022].

**Limitations** This chapter uses one of the simplest forgetting approach - directly resetting embeddings to random initialisation. Advanced techniques like gradually injecting noise could be explored. We focus on masked language modelling pretraining with language-specific tokenizers. Applying active forgetting to autoregressive LMs, other pretraining methods (e.g. DeBerta pretraining [He et al., 2021b,a]), and various tokenization strategies is promising future work. More broadly, current large language models need more plasticity to expand across tools, tasks, and domains. Our work takes an initial step, showing that directly resetting embeddings can significantly improve model plasticity. Further research on more sophisticated forgetting techniques during pretraining could unlock additional gains.

On the theory front, potential connections can be made between forgetting and meta-learning [Schaul and Schmidhuber, 2010, Thrun and Pratt, 2012, Andrychowicz et al., 2016, Finn et al., 2017] since both attempt to learn solutions that can quickly adapt themselves to new inputs. Another possible theoretical explanation for why active forgetting works so well might be related to the flatness of the solution in the loss landscape [Alabdulmohsin et al., 2021]. Flatter minima tend to enjoy better generalization [Liu et al., 2023b]. Thus, it might be worthwhile to study the flatness of the transformer body during the forgetting pretraining.

Beyond methodology, it would be valuable to more deeply investigate how this periodic resetting of embeddings affects the internal dynamics of the Transformer architecture itself. For instance, how does the reset influence attention patterns, layer activations, or representational drift across training epochs? Such analysis could shed light on whether active forgetting encourages more modular or adaptive representations. Additionally, while this work focuses on input embeddings, the same principle could be extended to other components such as attention heads or feedforward layers to improve plasticity further.

# 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. $\lambda$opt: 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.