Multitasking Models are Robust to Structural Failure: A Neural Model for Bilingual Cognitive Reserve

Giannis Daras*
The University of Texas at Austin
giannisdaras@utexas.edu

Negin Raoof*
The University of Texas at Austin
neginraoof@gmail.com

Zoi Gkalitsiou
The University of Texas at Austin
zoi.gkalitsiou@austin.utexas.edu

Alex Dimakis
The University of Texas at Austin
dimakis@austin.utexas.edu

Abstract

We find a surprising connection between multitask learning and robustness to neuron failures. Our experiments show that bilingual language models retain higher performance under various neuron perturbations, such as random deletions, magnitude pruning and weight noise compared to equivalent monolingual ones. We provide a theoretical justification of this robustness by mathematically analyzing linear representation learning and showing that multitasking creates more robust representations. Our analysis connects robustness to spectral properties of the learned representation and proves that multitasking leads to higher robustness for diverse task vectors.

1 Introduction

Converging evidence from cognitive science research indicates that bilingualism increases brain robustness by reducing the rate of cognitive decline due to aging [1, 2] and delaying the onset of symptoms of dementia [3, 4]. It appears that individuals who speak more than one language on a regular basis are able to maintain typical cognitive functioning despite neural degeneration. This mismatch between cognitive functioning and brain pathology is called cognitive reserve [5], and its underlying mechanisms are poorly understood and an active topic of investigation.

Inspired by this research, we study whether *artificial* neural networks are more robust when trained on multiple languages or multiple tasks. Our experiments demonstrate that training on multiple tasks indeed increases structural robustness. We train monolingual and bilingual GPT-2 models with the same architecture and dataset sizes. Initially, monolingual GPT-2 models are slightly outperforming the bilingual ones, but when we introduce structural noise (by randomly deleting neurons or adding noise to the weights) bilingual models degrade more gracefully and eventually outperform the monolingual models in the high-noise regime. For some amount of noise, bilingual models start outperforming the monolingual ones demonstrating a *cross-over* in performance due to their increased robustness. We observe this phenomenon for numerous models across three different types of corruption: Additive Gaussian noise to the weights, random weight pruning and magnitude-based weight pruning [6].

Our Contributions: We provide a theoretical justification of this phenomenon by mathematically analyzing linear multitask representation learning [7, 8]. Our analysis shows that introducing more diverse tasks creates ℓ_2 regularization in the linear task heads. Further, we formally connect the Euclidean norm of the learned representations to structural robustness under errors in the network

*equal contribution.

Preprint. Under review.

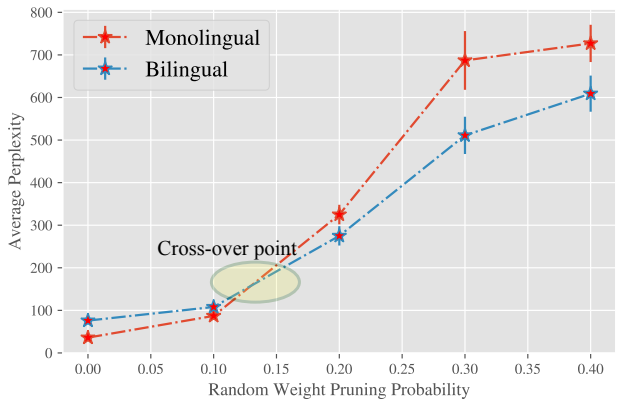

Figure 1: Performance of monolingual and bilingual GPT-2 models with the same architecture and training dataset size. We show the performance as we randomly erase weights. The x-axis indicates the probability of erasing an attention weight parameter (setting to it zero). The y-axis indicates the average perplexity over 20 runs with 95% confidence intervals. The bilingual model initially shows slightly worse performance, but as more weights are deleted, the monolingual model declines faster and worsens in the highly damaged regime. This indicates that the bilingual GPT-2 model is more robust to neuron weight erasures. We show similar results for several models and types of errors in our experimental section.

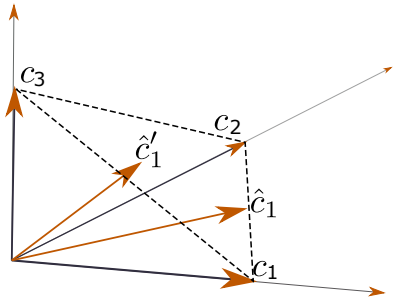

Figure 2: For two tasks, the best one dimensional approximation to c_1, c_2 is $\hat{c}_1 = [1/2, 1/2, 0]^T$ but the best one dimensional approximation to three tasks c_1, c_2, c_3 is $\hat{c}_1' = [1/3, 1/3, 1/3]^T$. Multi-tasking is creating ℓ_2 regularization since $||\hat{c}_1'||_2 < ||\hat{c}_1||_2$. It is important that the original task vectors c_1, c_2, c_3 are orthogonal i.e. diverse, since this creates regularization.

weights. Our main theorem is establishing that multitasking leads to higher robustness for linear representations when the task vectors are selected as random and independent Gaussian vectors under additive noise. Our results also establish that when the tasks are significantly overlapping, multitasking does not lead to higher robustness and hence task diversity is necessary.

We experimentally observe that multitasking increases structural robustness for numerous networks and multiple problems including MNIST, CIFAR10, Newsgroup20, GPT models and finetuned GPT models on GLUE tasks. We train networks under exactly comparable dataset and architecture conditions and show that models become more robust to structural failures as they are trained with more tasks. We experiment with three different types of structural failures and show robustness increases for all of them. We also experimentally observe that the addition of diverse tasks seems to regularize the model weights, as we predict in our theoretical analysis.

2 Theoretical Analysis

Building intuition. We start with a small numerical example to build intuition. Given a feature vector $x \in \mathbb{R}^d$ we compute a k dimensional linear representation Wx using a matrix $W \in \mathbb{R}^{k \times d}$. We choose W such that we best approximate a set of ground truth task vectors, $\{c_1, c_2, ..., c_T\}$, that lie in $\mathbb{R}^d$. The learned approximation is $\hat{c}_i = W^T \gamma_i$. Essentially, we use linear combinations of the columns

of W^T to approximate the task vectors. For simplicity, we assume that the columns of W^T are unit norm. We study the case where $k < T$, otherwise there are infinite solutions.

Assume we work in $d = 3$ dimensions with $T = 3$ total tasks, $c_1 = [1, 0, 0]^T, c_2 = [0, 1, 0]^T, c_3 = [0, 0, 1]^T$. Set our learned representation dimension to be $k = 1$ dimensional. When $T = 2$, using only the first two tasks c_1, c_2, an optimal solution is $W = \frac{1}{\sqrt{2}}[1, 1, 0]$. The corresponding linear head is now the scalar $\gamma_1 = \frac{1}{\sqrt{2}} = \gamma_2$ and the approximate vectors are $\hat{c}_1 = W^T\gamma_1 = [0.5, 0.5, 0]^T = \hat{c}_2$. Therefore the best one dimensional subspace to jointly approximate c_1, c_2 is the span of $W = \frac{1}{\sqrt{2}}[1, 1, 0]$. Now we introduce one more task and find the one dimensional subspace that best approximates c_1, c_2, c_3. That becomes $W' = \frac{1}{\sqrt{3}}[1, 1, 1]$ with linear heads $\gamma'_1 = \frac{1}{\sqrt{3}} = \gamma'_2 = \gamma'_3$. The approximate vectors now are $\hat{c}'_1 = (W')^T\gamma'_1 = [1/3, 1/3, 1/3]^T = \hat{c}'_2 = \hat{c}'_3$. Notice that $||\hat{c}'_i||^2 = 1/3$ for 3 tasks but $||\hat{c}_i||^2 = 1/2$ for two tasks. The point is that for *more tasks, the vector that jointly approximates all task vectors becomes shorter*. Equivalently, the ℓ_2 norm of the linear task heads *decreases* from $\gamma_i = \frac{1}{\sqrt{2}}$ to $\gamma'_i = \frac{1}{\sqrt{3}}$ as the tasks increased from two to three showing how multitasking creates regularization. A graphical representation of this example is given in Figure 2. It is important that the task vectors c_i are orthogonal, increasing the effective dimensionality of the problem. The intuition is that diverse tasks increase the effective dimension, making the best approximation vector shorter.

Our main theoretical result is showing that this phenomenon is quite general and makes multitasking lead to structural robustness. We connect the norm of the approximated task vectors with robustness to weight perturbations and show that for Gaussian, independent task vectors the average norm shrinks as more tasks are added. This is intuitive since high dimensional Gaussian vectors are near-orthogonal. Surprisingly, we empirically show that real task vectors for numerous problems also exhibit this behavior.

Analysis. We consider a neural network $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ and a collection of tasks $\{\mathcal{T}_1, ..., \mathcal{T}_T\}$. We are trying to learn $\theta, \gamma_i \in \mathbb{R}^k$ to solve the following optimization problem:

$$\text{argmin}_{\theta,\{\gamma_1,...,\gamma_T\}} \sum_{i=1}^{T} \mathbb{E}_{(x,y)\in\mathcal{T}_i} \mathcal{L}(\gamma_i^T f_\theta(x), y). \tag{1}$$

The neural network f_θ can be as simple as a single matrix $W : \mathbb{R}^d \to \mathbb{R}^k$. For linear networks, we consider the following dataset generation process: for task $\mathcal{T}_i$, we sample a Gaussian x and we generate its label y by taking the inner-product with a task vector c_i, i.e. $y = c_i^T x$ for task $\mathcal{T}_i$. Given infinite samples and MSE loss, the optimization problem of (1) is equivalent to the following problem.

Definition 2.1 (Optimization Problem)**.** Let $k < T < d$. We define the Factorized Best Rank-k approximation of a matrix $C \in \mathbb{R}^{d\times T}$ as the optimization problem:

$$W^*, \Gamma^* = \text{argmin}_{W\in\mathbb{R}^{k\times d},\Gamma\in\mathbb{R}^{k\times T}} \left|\left|W^T\Gamma - C\right|\right|_F^2. \tag{2}$$

We are interested in the case when the dimensionality of the representation k is smaller than the number of tasks T, otherwise the best Rank-k approximation of C is not unique.

The following Proposition states that in the considered setting, Problem 2 can be solved with SVD.

Proposition 2.2. *For any matrix $C \in \mathbb{R}^{d\times T}$ with distinct singular values, any solution of 2.1 satisfies:*

$$W^{*^T}\Gamma^* = U\Sigma_k V^T, \tag{3}$$

where $U\Sigma V^T$ is the SVD of C and Σ_k is the same as Σ except than the last $T - k$ diagonal entries that are zeroed out.

The fact that the Singular Value decomposition computes the best rank-k approximation to a matrix can be found in several textbooks e.g. Golub and Van Loan [9], Blum et al. [10].

This proposition establishes that $W^* = U^T$ and $\Gamma^* = \Sigma_k V^T$ is a valid solution of (2). Onwards, we will be calling this the SVD Solution.

Definition 2.3. We define the SVD solution of (2), to be:

$$W_{\text{SVD}} = U^T, \quad \Gamma_{\text{SVD}} = \Sigma_k V^T. \tag{4}$$

We note that if any multitask learning algorithm is used to obtain W^*, Γ^*, one can run Gram-Schmidt to make W^* orthonormal and hence obtain the factorization we use. It is important that W stays normalized and all scaling is pushed to Γ since to measure robustness to weight shifts, we are going to add noise to W only, and higher W scaling is equivalent to lower effective noise.

We study how the performance is affected when the representation network, f_θ, is corrupted.

Definition 2.4. For any sample x, the **Mean Squared Error (MSE)** for task i is defined to be the expected error between the model prediction under noise and the true value y. Namely,

$$\text{MSE}^i = \mathbb{E}_{\theta_c}\left[(\gamma_i^T f_{\theta_c}(x) - y)^2\right], \tag{5}$$

where f_{θ_c} is the model that emerges after corrupting f_θ.

This measures how well the model approximates the ground truth under the presence of noise and under the constraint of a joint representation for multiple tasks.

The simplest corruption process to study is adding noise to the representation matrix, i.e.

$$W_c = W + N, \quad N_{ij} \sim \mathcal{N}(0, \sigma^2), \text{ i.i.d} \tag{6}$$

Then, we denote the mean squared error for the task i with MSE^{i,σ^2} and the average mean squared error across the T tasks with $\overline{\text{MSE}}^{T,\sigma^2}$. We are now ready to introduce our results.

Theorem 2.5 (Mean Squared Error for Additive Noise)**.** *Let $C \in \mathbb{R}^{d\times T}$ be a matrix with distinct singular values $\sigma_1 > \sigma_2 > ... > \sigma_T$. Let W, Γ be the SVD solution of* (2)*. Under the Additive Noise Model defined in* (6)*, we have that:*

Average MSE without noise

$$\overline{\text{MSE}}^{T,\sigma^2} = \overline{\text{MSE}}^{T,0} + \frac{\sum_{i=1}^k \sigma_i(C)^2}{T} \cdot \sigma^2 \quad . \tag{7}$$

Average MSE under noise ↑ ↑ *Noise Variance*

As shown, the noisy MSE decomposes into the sum of the noiseless MSE plus the noise variance times a function that depends on the number of tasks:

$$R(T) = \frac{\sum_{i=1}^k \sigma_i(C)^2}{T}. \tag{8}$$

It is important to emphasize that as more tasks are added, the matrix C changes, but the interlacing theorem allows us to connect the singular values of smaller submatrices, as discussed in the Appendix. $R(T)$ is the robustness slope: if a model with T tasks has smaller slope, it will eventually outperform a model with, say $T-1$ tasks and larger slope, for sufficiently large noise. This is true even if the noiseless performance for the $T-1$-task model is better, indicating a cross-over in MSE. Therefore the key is understanding when the sum of the top k singular values of C scales sublinearly in T. This is not true for tasks that are aligned, but we can show it holds for independent Gaussian task vectors. We believe it holds for more general families of diverse task vectors and our experiments verify it also holds for numerous real task vectors learned from text and vision datasets.

Connection with l_2 regularization. For the SVD solution (see Definition 4), the sum of the top-k singular values squared is the squared Frobenius norm of Γ. Indeed, we have that $||\Gamma_{\text{SVD}}||_F^2 = ||\Sigma_k V^T||_F^2$. Since Σ_k is a diagonal matrix, each row of $\Sigma_k V^T$ is a rescaling of the corresponding row of V^T. Rows of V^T have norm 1, hence the i-th row of $\Sigma_k V^T$ will have norm σ_i. The Frobenius norm squared is just the sum of the squared norms of the rows. Hence, we get that

$$||\Gamma_{\text{SVD}}||_F^2 = \sum_{i=1}^k \sigma_i(C)^2. \tag{9}$$

Using this simple observation, we can get the following alternative expression of Theorem 2.5.

Corollary 2.6. *Let $C \in \mathbb{R}^{d\times T}$ be a matrix with distinct singular values. Let W, Γ be the SVD solution of* (2)*. Under the Additive Noise Model defined in* (6)*, we have that:*

$$\overline{\text{MSE}}^{T,\sigma^2} = \overline{\text{MSE}}^{T,0} + \frac{||\Gamma||_F^2}{T}\sigma^2 \quad . \tag{10}$$

Corollary 2.6 provides two important insights: i) the normalization with the number of tasks that appears in (7) is justified since the Frobenius norm of Γ grows with the number of task, ii) if we can prove that the slope (defined in Equation (8)) is dropping, then we are effectively proving that multitasking gives l_2 regularization as we showed in the toy introductory example. This also holds for the case of Gaussian, i.i.d. task vectors, as shown in the following theorem.

Theorem 2.7. *Let $C \in \mathbb{R}^{d\times T}$ be a random matrix with Gaussian, i.i.d. entries of variance $1/d$ and $d = \Omega(T^3)$. Let C_t, C_{t+1} be the matrices formed by selecting the first $t, (t+1)$ columns of C. Then, there is a noise level σ_{thres} such that with probability $\geqslant 1 - \exp\left(-\Omega\left(\sqrt{d}\right)\right)$, the SVD solutions (see (4)) of (2) (for C_t, C_{t+1} respectively), under the noise corruption model, satisfy:*

$$\overline{\text{MSE}}^{t+1,\sigma^2} < \overline{\text{MSE}}^{t,\sigma^2}, \quad \forall\, \sigma \geqslant \sigma_{\text{thres}}. \tag{11}$$

Remark 2.8. In words, this result shows that adding new tasks gives **provably** increased robustness to high noise corruption in the weights, when the task vectors are Gaussian.

Remark 2.9. Observe that the MSE under noise drops for *every single new task added*. The assumption $d = \Omega(T^3)$, can be relaxed to $d = \Omega(t^3)$, and we get increased robustness for the first t added tasks. Nevertheless, for most applications $d = \Omega(T^3)$ is a realistic assumption: Even for our smallest dataset MNIST $d = 728$, and we experiment with up to 10 tasks.

3 Experimental Evaluation

We divide the experimental section in two parts. In the first part, we add noise to the final linear representation layer of various networks and verify that our theoretical analysis agrees with experimentally observed multitasking robustness on real datasets (MNIST, CIFAR10, NewsGroup20). In the second part, we show that multitasking leads to robustness to general weight corruptions in any layer of a complex transformer. Specifically, we show that multilingual Language Models are more robust to weight shifts (across all the layers) compared to monolingual trained under the same setting. This is the first evidence of increased Cognitive Reserve in bilingual artificial neural networks.

Experiments with Linear Representation Layers. We perform experiments on three datasets (MNIST, CIFAR10, Newsgroup20) and two modalities (Vision and Language). The datasets normally involve one classification task each. We create multiple binary tasks by distinguishing between pairs of labels. For example, in CIFAR10, one task might be to distinguish between dogs and cats and another between airplanes and cars. We assign a value in $[0, 1]$ to each sample for each task to transform them to regression tasks (to match our theory). For example, if task i is to distinguish between dogs and cats, value 0 corresponds to dog and value 1 to cat.

The second issue is learning the task vectors from training data. For MNIST, we can simply learn a linear layer C with columns $\{c_1, ..., c_T\}$ such that: $c_i^T x \approx y$ for each task. For more complex datasets like CIFAR or Newsgroup20, linear networks have lower performance and hence it is less interesting to examine their robustness. Instead, we first use another network to extract representations $g_\theta(x)$ and then learn a linear layer acting on the encodings such that $c_i^T g_\theta(x) \approx y$. For CIFAR we used a pre-trained Resnet50 as the encoder while for NewsGroup, a pre-trained BERT [11]. We would like to point out that our theory is still valid for this case – this is equivalent to the linear layer C receiving inputs from a learned representation as opposed to the features directly. As the number of tasks increase, we reduce the number of training examples per task. We do this to make sure that the total training dataset size stays the same as the number of tasks increase.

Figure 3 shows how the average MSE behaves as noise increases for different number of tasks. Note that even though all models begin from roughly the same performance in the noiseless setting, the multitask models are much more robust to the corruption of their weights consistently among all the datasets and modalities. This is aligned with our theoretical analysis which predicts that the robustness slope (defined in Equation (8)) decreases with the number of tasks. We calculate robustness slopes for learned task vectors for real datasets and plot their decay in the Appendix, where we further include all the details of how these models were trained.

Experiments with Language Models. Our objective is to compare robustness to neural weight perturbations in monolingual and bilingual language models. We use the following perturbation

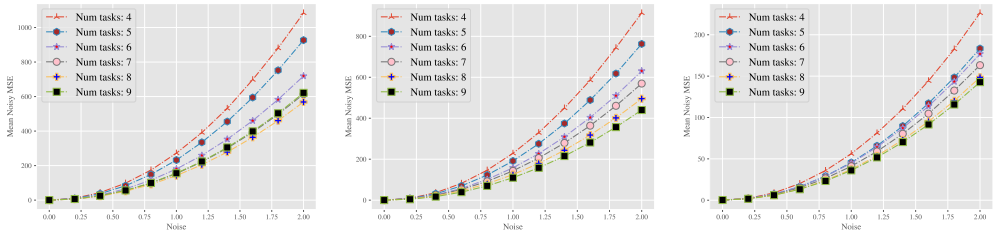

Figure 3: MSE of model (versus optimal task vector) as a function of noise added to the weights. From left to right: MNIST, CIFAR10, NewsGroup20. As shown for all these datasets, adding additional tasks is increasing the robustness of the model to weight noise.

models: 1) Random deletion of weight parameters: we zero-out p percent of the attention layer weights, 2) Magnitude pruning: we sort model attention weights by the magnitude and delete the smallest p percent of weights [6], 3) Random normal noise: we add zero-mean random Gaussian noise with standard deviation σ^2 to the attention weights.

On the selection of the linguistic pair, we selected Greek, a highly inflected language with very different morphology, syntax and phonology compared to English. It also uses a different script since Greek characters were not Romanized. This minimizes transfer between languages, something we wanted to avoid. In the Appendix, we present additional experiments for other Romance languages.

The dataset for the bilingual model is a concatenation of articles from English and Greek Wikipedia. To avoid the computational cost of training for a new language, we start from the pre-trained GPT-2 (small)[12] and we use the Language Model Recycling Technique, introduced in [13]. GPT-2 small is a transformer-based architecture for causal language modeling, with 12 attention blocks and 124M parameters. The tokenizer uses Byte Pair Encoding and has a vocabulary of $50,257$ tokens. For the bilingual model, we generate a new tokenizer, vocabulary and embedding layer without changing the architecture. We keep the vocabulary size the same, as changing the vocabulary size can affect the scale of the perplexity score for these models. Note that Wikipedia documents were not in the original training of GPT-2, but our monolingual baseline was subsequently finetuned on English Wikipedia. Details on all our training hyperparameters are included in the Appendix.

We measure the quality of generated text using perplexity. Our bilingual model achieves 89 perplexity on a randomly picked subset of the OSCAR [14] dataset and 76 perplexity on the English IMDB dataset [15]. Monolingual GPT-2 model achieves 36 perplexity on the IMDB dataset. In the Appendix we include generated text for both the models. Although the perplexity of the bilingual model does not match the pre-trained GPT-2, the generated text is of reasonable quality text in both languages.

Text Generation. Our first experiment is to compare the performance of both models under various parameter perturbations. First, we try deleting a random portion p (p from 0% to 40%) of attention layers' weight to observe and compare the trend of decay in text generation quality between the two models. We evaluate both models on the IMDB dataset. As the graph in Figure 1 shows, the monolingual model starts with text predictions closer to the source text, resulting in lower perplexity without noise. However, as we delete a more significant portion of weights, the bilingual model matches the performance of the monolingual one and eventually outperforms that.

Next, we try magnitude-based pruning of a portion of weights, p, to observe and compare the trend of decay in text generation quality between the two models. We sort the attention layer weights by the magnitude and set p percent of weights with the lowest magnitude to zero. Again, we use the IMDB dataset to evaluate models. The graphs in Figure 4 show that as the training process continues, the model achieves a lower perplexity. Moreover, pruning additional weights has a less substantial impact on the model's performance. This graph shows that training the pre-trained GPT-2 model for a few epochs on a bilingual dataset significantly improves robustness to weight perturbations.

In another experiment, we monitor the characteristics of the model parameters by observing the changes in the maximum singular value of the weight matrices throughout training process. We record the maximum singular value of attention layer weights. In this process, we use a pretrained GPT-2 model baseline, and train this model for 16k iterations on English text data from Wikipedia.

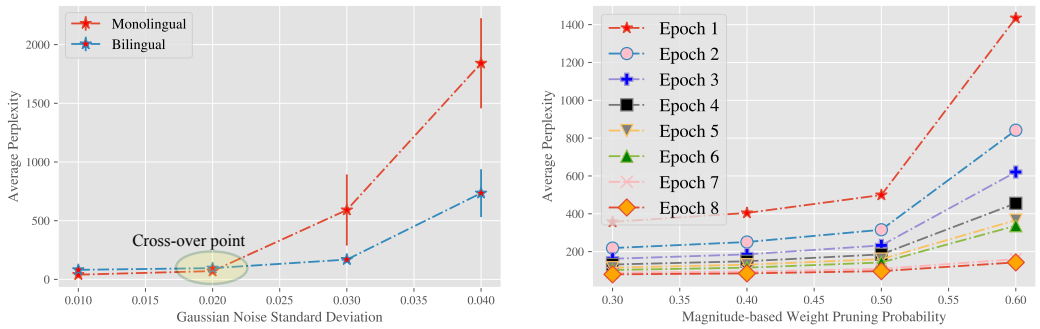

Figure 4: Robustness to magnitude-based weight pruning and additive Gaussian noise. When plotting perplexity under additive Gaussian noise, x-axis indicates the standard deviation of noise added to weights. Y-axis indicates the average perplexity over 20 runs with 95% confidence intervals. The second plot shows perplexity as we delete more weights based on magnitude, for the bilingual model at each epoch. X-axis indicates the probability of deleting sorted attention weight parameters. After only one epoch, the model shows higher sensitivity to weight perturbations. However, after eight epochs of training, it becomes more robust.

Resuming from this checkpoint, we train two new models: 1) We continue training model 1 on task 1 (English Wikipedia dataset) for 16k more iterations. 2) We train a second model on a different English dataset, the LAMBADA dataset [16], for 16k more iterations. Figure 5 indicates the results of this experiment by plotting maximum singular values of the first attention layer. As the Figure shows, training model on a new dataset (task 2) results in a faster decay of the max singular value.

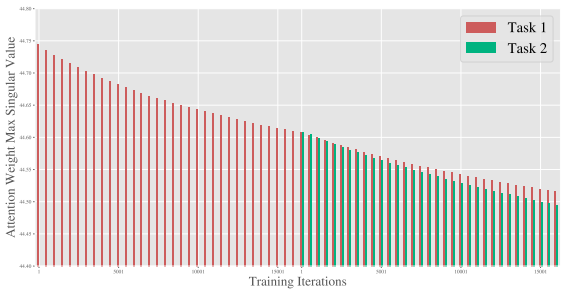

Figure 5: We show the effect of monolingual and bilingual training on the maximum singular value of attention weights. The red line shows the maximum singular value for a monolingual model trained on English Wikipedia for 32k iterations. The green line shows the maximum singular value if in the $16K$ iteration we switch to bilingual training. As shown, bilingual training leads to faster decay in the maximum singular value.

Text Classification. We conduct another set of experiments to observe the robustness of fine-tuned monolingual and bilingual GPT-2 models for text classification. In this section, we fine-tune both the monolingual and the bilingual GPT-2 models (previously trained) for downstream classification tasks using the GLUE benchmark [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28] to compare the robustness of models to weight perturbations. The two perturbation methods tested in this section are random weight deletion and random Gaussian noise added to attention weights. For each task, we fine-tune both models for ten epochs. When applying random pruning, the accuracy of each model is evaluated after deleting p percent of model weights, p ranging from 0% to 45%. When perturbing model weights by adding noise, we try various Gaussian noise distributions with standard deviations ranging from 0 to 0.09. Experiment results can be found in the Appendix section.

Random Pruning. We compare the classification accuracy between the fine-tuned model from the monolingual pre-trained network and the fine-tuned model using the bilingual network. Each element in attention parameters is pruned with probability p, where p ranges from .0 to .45. We evaluate the classification accuracy for the following GLUE tasks: CoLA, QQP, SST2, MRPC, QNLI, and RTE.

Task	Fine-tuned using	
	Monolingual ckpt	Bilingual ckpt
SST2	70567.875	60663.121
QQP	70608.195	60649.586
MRPC	70498.953	60590.769
RTE	70508.968	60590.765
CoLA	70519.781	60600.933

Table 1: We compute the sum of the squares of the weights of an attention layer for monolingual and bilingual models. The latter have smaller magnitudes, indicating that multitasking induces weight regularization.

Pruning Probability	QQP		SST2		COLA		MRPC		RTE	
	m.	b.	m.	b.	m.	b.	m.	b.	m.	b.
0.00	0.876	0.843	0.908	0.862	0.437	0.218	0.828	0.774	0.646	0.595
0.05	0.873	0.842	0.909	0.866	0.425	0.203	0.804	0.769	0.640	0.589
0.10	0.867	0.833	0.899	0.868	0.403	0.204	0.730	**0.744**	0.603	0.575
0.15	0.848	0.819	0.871	0.866	0.366	0.185	0.619	**0.730**	0.600	0.562
0.20	0.804	0.786	0.836	**0.859**	0.326	0.179	0.416	**0.663**	0.561	0.553
0.25	0.711	**0.732**	0.806	**0.847**	0.267	0.159	0.377	**0.653**	0.543	0.546
0.30	0.656	**0.678**	0.760	**0.828**	0.216	0.137	0.320	**0.504**	0.537	0.536
0.35	0.638	**0.674**	0.714	**0.815**	0.153	0.092	0.317	**0.420**	0.522	0.494
0.40	0.632	**0.655**	0.683	**0.793**	0.097	0.058	0.316	**0.328**	0.521	0.488
0.45	0.632	**0.636**	0.651	**0.773**	0.060	0.042	0.316	**0.328**	0.525	0.485

Table 2: Performance under a range of random pruning probabilities for various GLUE tasks. Columns labeled with "m" determine classification accuracy of monolingual models and columns labeled as "b" determine accuracy of bilingual. CoLA is evaluated using Matthew's Correlation and other tasks are evaluated by accuracy.

We expect the accuracy of both models to decay as we prune a more considerable number of parameters. The monolingual model shows a faster decay in almost all tasks. For some tasks such as SST2, QQP, and MRPC, we observe that the bilingual model starts with lower accuracy, and its performance exceeds the monolingual model as we prune $\approx 5\%$ to $\approx 25\%$ of parameters. A detailed set of results in Table 2 show models' average prediction accuracy on the GLUE benchmark.

Random Noise. We also experiment with adding Gaussian noise to the weights. We vary the noise standard deviation from .0 to 0.09. We evaluate the classification accuracy for the same tasks. When no noise is added to model parameters, the monolingual model performs slightly better for tasks like QQP and SST2. As we increase the noise, the accuracies of both models drop with almost identical rates. However, both graphs illustrate a cross-over point after which the bilingual model outperforms the monolingual. The bilingual model achieves significantly higher accuracy in the MRPC task when the standard deviation is greater than ≈ 0.03. For CoLA and RTE, the monolingual model maintains maintains higher performance regardless of the noise level. A detailed set of results in the Appendix section shows models' average prediction accuracy on the GLUE benchmark.

4 Related Work

Cognitive Reserve and Bilingualism. Our work is inspired by Cognitive Science and evidence of Cognitive Reserve in bilinguals. One implication of our theory is that multitasking leads to smaller weights on average. This could be related to studies performed in healthy older adults that indicate that despite overall less gray matter volume and poorer white matter integrity (i.e., poorer structural brain connectivity), older healthy bilinguals perform equally well or outperform monolinguals in several cognitive tasks [1, 2].

We would like to emphasize that our research is solely on *artificial networks* which have huge differences to biological neurons. No definite extrapolations should be made to Cognitive Neuroscience without further work. Nonetheless, we show that there is a simple mathematical abstraction that seems to align with the significantly more complex phenomena observed in bilingual cognitive reserve.

Multitask Learning. The most closely related work is by Mao et al. [29] which shows that multitask learning increases *adversarial* robustness. The intuition behind their proof is that, with task diversity, the gradient of the loss with respect to the wrong label is small as orthogonal tasks make gradients

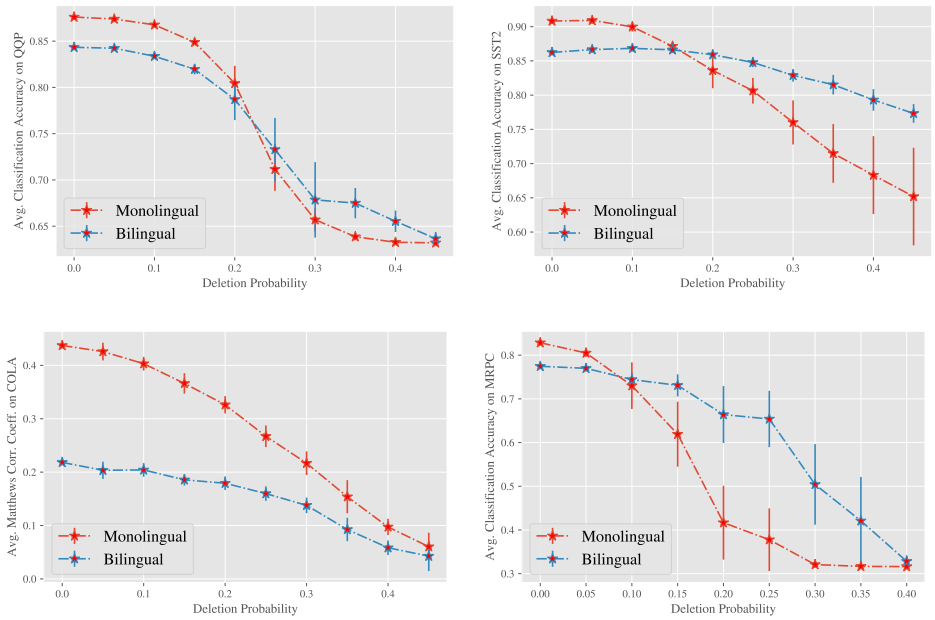

Figure 6: Performance comparison in GLUE tasks: QQP, SST2, CoLA, and MRPC under random erasures. QQP: Monolingual drops lower than the bilingual model after $\approx 25\%$ of the parameters are deleted. SST2: Monolingual drops with a faster rate, falling behind the bilingual after deleting $\approx 15\%$ of the parameters. CoLA: Both models reach ≈ 0 MCC (random prediction) with $\approx 45\%$ of parameters pruned. MRPC: The accuracy of the monolingual degrades at a faster rate as pruning probability increases higher than $\approx 10\%$.

that cancel out. Wu et al. [30] establishes a connection between robustness to weight perturbations and robustness to adversarial attacks. Our work is related but different since it directly establishes a connection between structural robustness and multitasking and shows a cross-over in performance across various domains and tasks. Our theoretical analysis is also completely different compared to prior works. More information on multitask learning can be found in Mao et al. [29] and Ghamizi et al. [31].

Many studies on network compression and the Lottery Ticket Hypothesis are related to our Magnitude Pruning experiments. LeCun et al. [32], Han et al. [6] find that selectively pruned networks can be trained from randomly initialized weights to match the performance of the original network. Frankle and Carbin [33] introduces the hypothesis that randomly initialized neural networks contain a very sparse sub-network that, if initialized correctly, can achieve the accuracy of the original model. Chen et al. [34] studies this in continual learning and examines various pruning methods.

5 Conclusions

We demonstrated a connection between multitask learning and robustness to structural failures for artificial neural networks. For linear representation learning we obtained a characterization of robustness through the spectrum of the task matrix. We showed that robustness comes from diverse tasks which imply a bounded spectral norm for C. One limitation of our theoretical work is that we did not analyze learning algorithms but directly used the SVD solution. It would be interesting to see if gradient descent introduces further regularization or other effects, especially in the non-linear case.

Experimentally, we observed increased robustness for both linguistic and non-linguistic tasks. More complex settings like multi-lingual models, cross-language transfer and their interactions remain to be explored. Finally, it remains open if bilingualism and cognitive reserve in humans can indeed be connected to our framework. It would be fascinating if neuroimaging techniques can measure any form of anatomical or functional regularization that bilingualism could be creating in humans.

6 Acknowledgments

This research has been supported by NSF Grants CCF 1763702, 1934932, AF 1901281, 2008710, 2019844 the NSF IFML 2019844 award as well as research gifts by Western Digital, WNCG and MLL, computing resources from TACC and the Archie Straiton Fellowship.

A Proofs

Theorem 2.5. *Let $C \in \mathbb{R}^{d\times T}$ be a matrix with distinct singular values $\sigma_1 > \sigma_2 > ... > \sigma_T$. Let W, Γ be the SVD solution of* (2)*. Under the Additive noise model defined in 6,*

$$\overline{\mathrm{MSE}}^{T,\sigma^2} = \overline{\mathrm{MSE}}^{T,0} + \frac{\sum_{i=1}^{k}\sigma_i^2(C)}{T}\sigma^2 \ . \tag{12}$$

Proof.

$$\mathrm{MSE}^{i,\sigma^2} = \mathbb{E}\left[\left(c_i^T x - \gamma_i^T W_c x\right)^2\right] = x^T c_i c_i^T x - 2c_i^T x\gamma_i^T\mathbb{E}[W_c]x + x^T\mathbb{E}[W_c^T\gamma_i\gamma_i^T W_c]x \tag{13}$$

$$= x^T c_i c_i^T x - 2c_i^T x\gamma_i^T W x + x^T\mathbb{E}[W_c^T\gamma_i\gamma_i^T W_c]x \tag{14}$$

$$= x^T c_i c_i^T x - 2c_i^T x\gamma_i^T W x + x^T\mathbb{E}[(W^T + N^T)\gamma_i\gamma_i^T(W+N)]x \tag{15}$$

$$= x^T c_i c_i^T x - 2c_i^T x\gamma_i^T W x + x^T W^T\gamma_i\gamma_i^T W x + x^T\mathbb{E}[N^T\gamma_i\gamma_i^T N]x \ . \tag{16}$$

Now, observe that:

$$\sum_{i=1}^{T}\mathrm{MSE}^{i,\sigma^2} = \left(\sum_{i=1}^{T} x^T c_i c_i^T x - 2c_i^T x\gamma_i^T W x + x^T W^T\gamma_i\gamma_i^T W x\right) + x^T\mathbb{E}[N^T\gamma_i\gamma_i^T N]x \tag{17}$$

$$\overline{\mathrm{MSE}}^{T,\sigma^2} = \overline{\mathrm{MSE}}^{T,0} + \frac{x^T\mathbb{E}\left[N^T\left(\sum_{i=1}^{T}\gamma_i\gamma_i^T\right)N\right]x}{T} \ . \tag{18}$$

Observe that

$$\mathbb{E}\left[N^T\left(\sum_{i=1}^{T}\gamma_i\gamma_i^T\right)N\right] = \sigma^2\mathrm{tr}\left(\sum_{i=1}^{T}\gamma_i\gamma_i^T\right)I_d \ . \tag{19}$$

For any unit-norm x:

$$x^T\mathbb{E}\left[N^T\left(\sum_{i=1}^{T}\gamma_i\gamma_i^T\right)N\right]x = \sigma^2\mathrm{tr}\left(\sum_{i=1}^{T}\gamma_i\gamma_i^T\right) \ . \tag{20}$$

Now for the SVD solution, we know that $\Gamma = \Sigma_k V^T$ and $\{\gamma_i\}$ are the columns of Γ. Hence,

$$\sum_{i=1}^{T}\gamma_i\gamma_i^T = \Sigma_k^2 V^T V = \Sigma_k^2. \tag{21}$$

Then,

$$\overline{\mathrm{MSE}}^{T,\sigma^2} = \overline{\mathrm{MSE}}^{T,0} + \sigma^2\frac{\sum_{i=1}^{k}\sigma_i(C)^2}{T} \ . \tag{22}$$

□

We are going to use the following result due to Ledoux [35].

Lemma A.1 (Ledoux [35])**.** *Let $C_T : \mathbb{R}^{d\times T}$ be a random matrix whose entries are i.i.d. Gaussian with variance $1/d$. Let C_K be the random matrix that submatrix of C that consists of the first K columns of C_T. Then,*

$$\Pr\left[\sigma_{\max}(C_K) \geqslant 1 + \sqrt{K/d} + o(1) + \alpha\right] \leqslant \exp(-d\alpha^2/2) \tag{23}$$

and

$$\Pr\left[\sigma_{\min}(C_K) \geqslant 1 - \sqrt{K/d} + o(1) - \alpha\right] \leqslant \exp(-d\alpha^2/2) \ , \tag{24}$$

where $o(1)$ is a small-term that tends to 0 as $d \to \infty$.

Theorem 2.7. *Let $C \in \mathbb{R}^{d\times T}$ be a random matrix with Gaussian, i.i.d. entries of variance $1/d$ and $d = \Omega(T^3)$. Let C_t, C_{t+1} be the matrices that are formed by selecting the first $t, (t+1)$ columns of C respectively. Then, there is a noise level σ_{thres} such that with probability $\geqslant 1 - \exp\left(-\Omega\left(\sqrt{d}\right)\right)$, the SVD solutions (see* (4)*) of* (2) *(for C_t, C_{t+1} respectively), under the noise corruption model, satisfy:*

$$\overline{\text{MSE}}^{t+1,\sigma^2} < \overline{\text{MSE}}^{t,\sigma^2} \quad , \tag{25}$$

$\forall \sigma \geqslant \sigma_{\text{thres}}$.

Proof. From Theorem 2.5, we have that:

$$\overline{\text{MSE}}^{t+1,\sigma^2} = \overline{\text{MSE}}^{t+1,0} + \frac{\sum_{i=1}^{k}\sigma_i^2(C_{t+1})}{t}\sigma^2, \qquad \overline{\text{MSE}}^{t,\sigma^2} = \overline{\text{MSE}}^{t,0} + \frac{\sum_{i=1}^{k}\sigma_i^2(C_t)}{t}\sigma^2 \ . \tag{26}$$

To prove the desired thing, we just need to show that $\overline{\text{MSE}}^{t+1,\sigma^2}$ has a smaller co-efficient for the term σ^2, because for large enough σ, eventually this term will dominate the sum. Hence, we need to show that:

$$\frac{\sum_{i=1}^{k}\sigma_i^2(C_t)}{t} \geqslant \frac{\sum_{i=1}^{k}\sigma_i^2(C_{t+1})}{t+1} \ . \tag{27}$$

Since C_t is a submatrix of C_T, from the Eigenvalue Interlacing Theorem we know that $\sum_{i=1}^{k}\sigma_i^2(C_t) \leqslant \sum_{i=1}^{k}\sigma_i^2(C_{t+1})$. However, the difference of the two sums is upper-bounded. Using Lemma A.2, we get that:

$$\sum_{i=1}^{k}\sigma_i^2(C_{t+1}) = \sigma_1^2(C_{t+1}) + \sum_{i=2}^{k}\sigma_i^2(C_{t+1}) \tag{28}$$

$$\leqslant \sigma_1^2(C_{t+1}) + \sum_{i=1}^{k-1}\sigma_i^2(C_t) \tag{29}$$

$$= \sigma_1^2(C_{t+1}) - \sigma_k^2(C_t) + \sum_{i=1}^{k}\sigma_i^2(C_t). \tag{30}$$

It suffices to show that:

$$\frac{\sum_{i=1}^{k}\sigma_i^2(C_t)}{t} \geqslant \frac{\sigma_1^2(C_{t+1}) - \sigma_k^2(C_t) + \sum_{i=1}^{k}\sigma_i^2(C_t)}{t+1} \iff \tag{31}$$

$$\sum_{i=1}^{k}\sigma_i^2(C_t) \geqslant t\left(\sigma_1^2(C_{t+1}) - \sigma_k^2(C_t)\right). \tag{32}$$

Trivially, $\sum_{i=1}^{k}\sigma_i^2(C_t) \geqslant k\sigma_k^2(C_t)$. Hence, it is enough to show that:

$$\sigma_k^2(C_t) \geqslant \frac{t}{k}\left(\sigma_1^2(C_{t+1}) - \sigma_k^2(C_t)\right). \tag{33}$$

We will now bound the difference of the first and k-th singular values.

From Lemma A.1, we have that:

$$\Pr\left[\sigma_1(C_{t+1}) \geqslant 1 + o(1) + \sqrt{\frac{t+1}{d}} + \alpha\right] \leqslant \exp\left(-d\alpha^2/2\right) \tag{34}$$

and

$$\Pr\left[\sigma_k(C_t) \leqslant 1 + o(1) - \sqrt{\frac{t}{d}} - \alpha\right] \leqslant \exp(-d\alpha^2/2). \tag{35}$$

By union bound, with probability $\geqslant 1 - 2\exp(-d\alpha^2/2)$, we have that:

$$\sigma_1^2(C_{t+1}) - \sigma_k^2(C_t) \leqslant \left(1 + o(1) + \sqrt{\frac{t+1}{d}} + \alpha\right)^2 - \left(1 + o(1) - \sqrt{\frac{t}{d}} - \alpha\right)^2 \tag{36}$$

$$= 2(1 + o(1))\left(\sqrt{\frac{t+1}{d}} + \alpha\right)\left(\sqrt{\frac{t}{d}} + \alpha\right) + \left(\sqrt{\frac{t+1}{d}} + \alpha\right)^2 - \left(\sqrt{\frac{t}{d}} + \alpha\right)^2 \tag{37}$$

$$\leqslant 5\left(\sqrt{\frac{t+1}{d}} + \alpha\right)^2. \tag{38}$$

We choose $\alpha = \left(\frac{t+1}{d}\right)^{1/4}$. Since, $t < T < d$, we have that with probability $\geqslant 1 - \exp\left(-\Omega\left(\sqrt{d}\right)\right)$,

$$\sigma_1^2(C_{t+1}) - \sigma_k^2(C_t) \leqslant 20\sqrt{\frac{t+1}{d}}, \quad \sigma_k(C_t) \leqslant 1 + o(1) - 2\sqrt{\frac{t+1}{d}}. \tag{39}$$

Going back to Eq. 33, it suffices to show that:

$$1 + o(1) - 2\sqrt{\frac{t+1}{d}} \geqslant \frac{20t}{k}\sqrt{\frac{t+1}{d}} \iff 1 + o(1) \geqslant \sqrt{\frac{t+1}{d}}\left(\frac{20t}{k} + 2\right). \tag{40}$$

Since $t < T$, this is true for $d = \Omega(T^3)$.

□

Lemma A.2. *Let C be a matrix $\in \mathbb{R}^{d\times T}$ and $c_{T+1} \in \mathbb{R}^d$. Let also $\mathrm{C_{new}} = \begin{bmatrix} C & c_{T+1}\end{bmatrix} \in \mathbb{R}^{d\times T+1}$. Denote with $\sigma_i(C)$ the i-th singular value of C, sorted from the largest to the smallest. Then,*

$$\sigma_{i+1}(\mathrm{C_{new}}) \leqslant \sigma_i(C) \leqslant \sigma_i(\mathrm{C_{new}}), \quad \forall i \in \{1, ..., T\} \tag{41}$$

Proof. We have that:

$$\mathrm{C_{new}}^T\mathrm{C_{new}} = \begin{bmatrix} C^T \\ c_{T+1}^T\end{bmatrix} \cdot \begin{bmatrix} C & c_{T+1}\end{bmatrix} = \begin{bmatrix} C^TC & C^Tc_{T+1} \\ c_{T+1}^TC & c_{T+1}^Tc_{T+1}\end{bmatrix}. \tag{42}$$

Observe that $\mathrm{C_{new}}^T\mathrm{C_{new}}$ is a symmetric matrix and C^TC is a principal submatrix. Hence, from the Eigenvalue Interlacing Theorem, we have that:

$$\lambda_{i+1}(\mathrm{C_{new}}^T\mathrm{C_{new}}) \leqslant \lambda_i(C^TC) \leqslant \lambda_i(\mathrm{C_{new}}^T\mathrm{C_{new}}), \tag{43}$$

where $\lambda_i(A)$ is the i-th eigenvalue of A, sorted from the largest to the smallest. To finish the proof, we note that for any matrix A, $\sigma_i(A) = \sqrt{\lambda_i(A^TA)}$. □

B Additional Results

In this section, we include additional results that further support the findings of the main paper.

Robustness slope Recall Theorem 2.5 of the paper.

$$\overline{\text{MSE}}^{T,\sigma^2} = \overline{\text{MSE}}^{T,0} + \frac{\sum_{i=1}^{k} \sigma_i(C)^2}{T} \cdot \sigma^2 \ . \tag{44}$$

Average MSE without noise

Average MSE under noise

Robustness slope

Noise Variance

This theoretical finding implies that the cross-over phenomenon that we observe in our experiments (at least for the linear case), stems from a lower Robustness Slope in the multitask models. Figure 3 shows that the MSE under noise is lower for models that are trained to do more tasks. In Figure 7 of this Appendix, we show that indeed this is due to a decrease in the robustness slope. Across three different datasets, MNIST, CIFAR10, NewsGroup20, we see that increasing the number of tasks leads to a decrease in the robustness slope. We note that this does not necessarily mean a monotonic decrease in the MSE under noise. Since the total dataset size and the parameter k stay the same, increasing the number of tasks usually leads to increased noiseless MSE. However, under the presense of noise, our theory predicts (and our experiments confirm) that eventually the multitask model will reach superior performance.

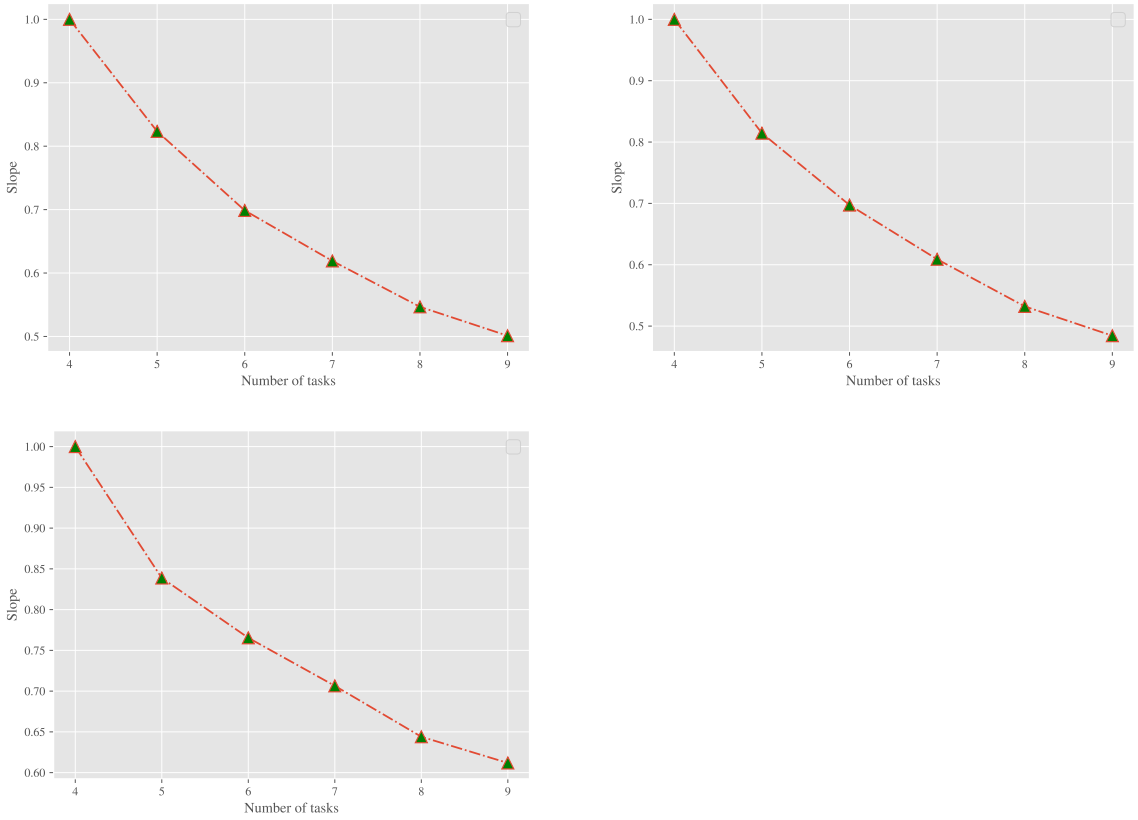

Figure 7: Slope as a function of the number of tasks for different datasets. (1, 1): MNIST, (1, 2): CIFAR10, (2, 1): NewsGroup20. As shown, adding more tasks decreases the robustness slope which leads to an increase in robustness (see Theorem 2.5).

Experiments on other languages For our experiments on multilingual generative models, we decided to use Greek and English because we were looking for a linguistic pair with different morphology, syntax and phonology. This is inspired by our theory on linear models that shows that diversity in the tasks (as we have for the Gaussian task vectors) leads to a sublinear increase in the sum of the top-k singular values of the task matrix and hence an increase in robustness. For completeness, we include here experiments on a different linguistic pair, English and Spanish. English and Spanish much closer linguistically and also share the Latin alphabet, so we expect bigger transfer and smaller robustness benefit in this linguistic pair.

We compare a monolingual English model (finetuned on English Wikipedia) with a bilingual, English and Spanish, model. The bilingual model is finetuned on a concatenation of English and Spanish Wikipedia. We make sure that the total dataset size is the same for the monolingual and the bilingual model, i.e. the bilingual model is exposed to half English data compared to the monolingual. This ensures that any benefits in terms of robustness are not coming from exposure to more data. We present results on random deletions in Figure 8 of the Appendix – this Figure is similar to Figure 1 of the paper, but instead of having English and Greek, we have English and Spanish. As shown in Figure 8, even though the two models are starting from roughly the same perplexity, the bilingual model exhibits higher structural robustness in the presence of weight deletions. This is consistent with the results we showed across this paper and indicates that the increased robustness is not specific to the choice of the linguistic pair.

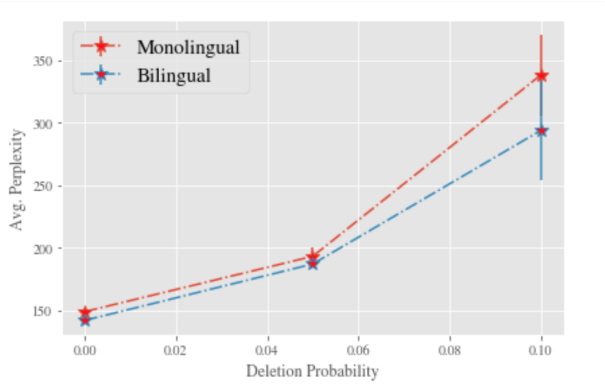

Figure 8: Performance of monolingual (English) and bilingual (English/Spanish) GPT-2 models with the same architecture and training dataset size. The x-axis indicates the probability of erasing an attention weight parameter (setting to it zero). The y-axis indicates the average perplexity over 20 runs. Models have a close initial accuracy. Perplexity increases (showing lower accuracy) as weight deletion probability is increased, though bilingual model perplexity rises at a slower rate.

Notice that the gap in the performance is smaller compared to the one presented in Figure 1. This is aligned with our theory for linear models that predicts that the benefits of multitasking for robustness are more evident for more diverse tasks. Since English and Spanish are linguistically closer, compared to English and Greek, our intuition is that the difference in robustness is going to be smaller and this is also confirmed by this experiment. An interesting future direction is to study this robustness benefit for multiple linguistic pairs or multi-lingual models. However, this study requires massive computational resources. Similarly, it would be interesting to study how the robustness gap in bilingual models scales as the datasets scale, but this also requires training multiple pairs of GPT models to comparable accuracy, and requires computational resources that were not available to us. We hope that future research is going to shed more light into these exciting directions.

Experiments with different corruption mechanisms. In the main paper, we primarily presented results with random deletions of neurons as our corruption model for the language modeling experiments. We include results for additive Gaussian noise for GPT-2 (monolingual and bilingual). We choose to present additional results with this noise model since it is the one analyzed by our theory. Table 3 summarizes how the performance of GPT-2 (monolingual and bilingual) changes when we add different amount of noise to the weights. We evaluate this performance on downstream tasks from the GLUE paper. Figure 9 visualizes the decrease of performance as the magnitude of the noise rises for different number of tasks. The results are similar with the results presented in the main paper for random deletions. In QQP, the monolingual model performs better without perturbations. Both models decay with a close rate. The monolingual model outperforms in SST-2 with no perturbations. Both models decay with a close rate. For CoLA, the monolingual model maintains a significantly better performance regardless of the noise level. Finally, for MRPC we see that although the bilingual model shows a weaker classification accuracy with no noise, it outperforms the monolingual model for noise levels higher than 0.035. These results complement Figure 4 of the main paper that shows robustness of GPT-2 to additive Gaussian noise for the task of language modeling.

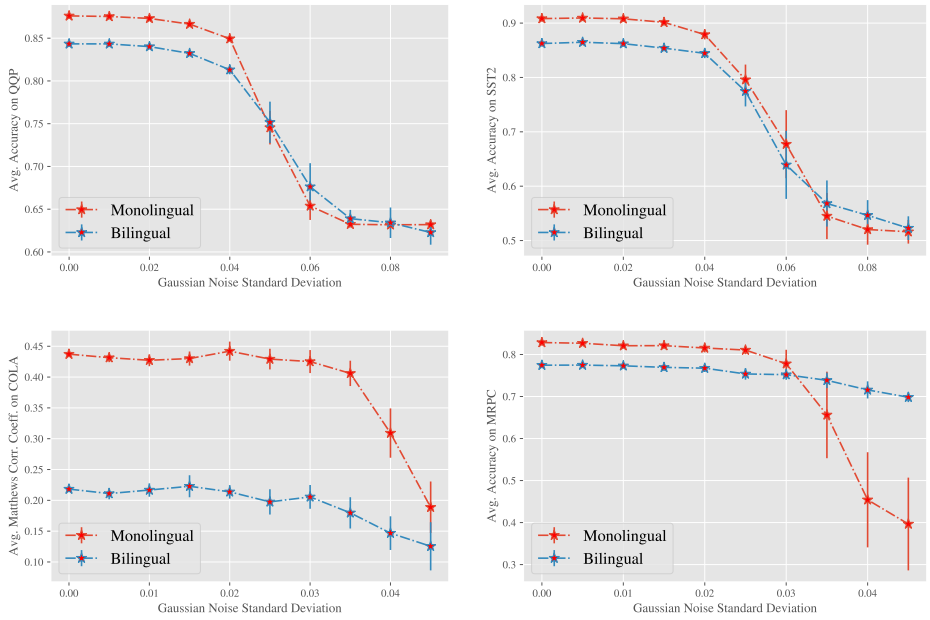

Figure 9: Performance comparison in GLUE tasks: QQP, SST2, CoLA, and MRPC under Gaussian noise. QQP: The monolingual model performs better without perturbations. Both models decay with a close rate. SST2: The monolingual model outperforms with no perturbations. Both models decay with a close rate. CoLA: The monolingual model maintains a significantly better performance regardless of the noise level. MRPC: Although the bilingual model shows a weaker classification accuracy with no noise, it outperforms the monolingual model for noise levels higher than 0.035.

Gaussian std.	QQP		SST2		COLA		MRPC		RTE	
	m.	b.	m.	b.	m.	b.	m.	b.	m.	b.
0.00	0.876	0.843	0.908	0.862	0.437	0.218	0.828	0.774	0.646	0.595
0.01	0.875	0.843	0.909	0.864	0.432	0.216	0.827	0.772	0.639	0.597
0.02	0.873	0.840	0.907	0.862	0.444	0.213	0.816	0.760	0.641	0.591
0.03	0.866	0.832	0.901	0.853	0.440	0.205	0.776	0.749	0.643	0.590
0.04	0.849	0.813	0.878	0.844	0.316	0.146	0.494	**0.711**	0.634	0.589
0.05	0.745	**0.751**	0.795	0.774	0.088	0.074	0.326	**0.673**	0.622	0.577
0.06	0.653	**0.676**	0.677	0.639	-0.002	0.004	0.316	**0.610**	0.602	0.563
0.07	0.632	**0.638**	0.545	**0.568**	0.019	0.002	0.316	**0.577**	0.585	0.562
0.08	0.631	**0.634**	0.520	**0.546**	-0.006	-0.006	0.316	**0.465**	0.539	0.546
0.09	0.631	0.622	0.516	**0.522**	-0.014	0.002	0.316	**0.451**	0.536	0.528

Table 3: Performance on GLUE when adding Gaussian noise. Columns labeled with ”m” determine classification accuracy of monolingual models and columns labeled as ”b” correspond to bilingual models. CoLA is evaluated using Matthew’s Correlation and other tasks are evaluated by accuracy.

[the company produces] video games, television programs, and online services. the company is headquartered in new york city and is the world's second largest entertainment company in terms of revenue, after comcast. disney was founded on october 16, 1923, by brothers walt disney and roy o'brien, jr.

[η εταιρεία παράγει] και διανέμεται στο χρηματιστήριο αθηνών, το οποίο παρέχει υπηρεσίες για τη διαχείριση των υπηρεσιών της εταιρείας. η εταιρεία είναι επίσης υπεύθυνη για την προώθηση του διαδικτύου σε συνεργασία με άλλες εταιρείες που δραστηριοποιούνται στην ευρωπαϊκή ραδιοτηλεοπτική ένωση

Table 4: Sample text generated by the bilingual GPT-2 model. Text in the brackets is the starting prompt provided for model.

C Training Details

Using the GPT-2 small model as our baseline, we fine-tuned a monolingual (English) model and a bilingual (English and Greek) model on Wikipedia text data. With set the vocabulary size to 50257 tokens. In both training processes, we set the initial learning rate to 3e-4 and configured a cosine learning rate scheduler with 150 warmup steps, setting AdamW optimizer weight decay to 0.01. We trained each model for eight epochs, using 4 NVIDIA Quadro RTX 5000 GPUs. Training took approximately 10 hours per epoch.

To fine-tune another bilingual model on English and Spanish data, we fine-tuned a monolingual model and a bilingual model on Wikipedia text data. With a vocabulary size of 50257 tokens, the monolingual model was fine-tuned on 800,000 English articles. The bilingual model was fined tuned on a mix of 400,000 Spanish and 400,000 English articles, using the same vocabulary size of 50257. Like the previous experiment, we set the initial learning rate to 3e-4 and configured a cosine learning rate scheduler with 150 warmup steps, setting AdamW optimizer weight decay to 0.01.

We further tuned bilingual and monolingual models for the text classification experiments using GLUE datasets. For these experiments, we used the AdamW optimizer with a learning rate of 2e-5, and epsilon at 1e-8. We used a linear scheduler with no warmup steps and trained models for more than ten epochs.

For the experiments on the linear representation layer, we used Adam optimizer with weight decay $1e-4$. We trained all the models with a batch size of 128 and to a maximum of 50 epochs. To emulate multiple tasks, we selected different subsets of the classes. We experiments with having class overlaps (e.g. for MNIST one task might have been predicting 0 vs 1 and some other task predicting 1 vs 2) and without class overlaps (e.g. predicting 0 vs 1 and 2 vs 3). We noticed bigger robustness benefits when there was no class overlap something that is consistent with our theoretical analysis that implies that diversity in the tasks is needed. In terms of corruptions, we also did preliminary experiments on random deletions and we saw similar results. The interested reader might use the released code to perform other types of weights corruptions and see how this affects robustness trends. For all our experiments, we fix the representation dimension to $k = 4$, which is also why we show the robustness slope from $k = 4$ onwards on Figure 7. Training time of the linear experiments depends on the dataset size: it took us roughly 1 hour on CIFAR-10 and 3 hours on NewsGroup20.

D Things that did not work

In the early stages of the project, we attempted to train a bilingual model from scratch, instead of using the recycling technique [13]. The dataset for the Greek model consists of roughly 2GB of text from Wikipedia. With such limited amount of data, we found it impossible to train a bilingual model that reaches a reasonable perplexity. Note that GPT-2 was trained on ≈ 40GB of text, i.e. on a $\approx 20\times$ bigger dataset. We found that the recycling technique [13] enables learning with much smaller datasets (on top of the computational benefits it offers).

E Limitations and Ethical Considerations

Limitations Even though the models we train can produce text of reasonable quality (e.g. see Table 4), they do not perform on par with state-of-the-art generative networks. There are many reasons for that, e.g. we do not have the computational resources to train bigger networks and the dataset size is small. Nevertheless, the goal of this paper is not to advance the state-of-the-art in text-generation but to shed light on how multitasking is related to robustness.

The motivation of this paper was a theory from Cognitive Science about increased robustness in bilingual speakers. We see that bilingual artificial networks are also more robust compared to monolingual models trained under the same setting. However, it is important to state that no definite extrapolations should be made to Cognitive Neuroscience without significantly much work. Our models of corruptions happening to the neural network's weights are chosen primarily for simplicity in the implementation and in the analysis. There is no evidence that brain pathologies have any resemblance to the models of corruption analyzed in this work for artificial neural networks.

Finally, our theory did not analyze the learning dynamics for approximating the task vectors. Instead, it used the SVD solution. Different choices of learning algorithms might lead to different behaviors regarding robustness. For example, for the linear case we showed that multitasking creates weight regularization. Higher explicit weight regularization (e.g. with high weight decay) might help the single task model decrease the robustness gap with the multitask networks. It would also be interesting to explore how the theory can be generalized to the non-linear case.

Ethical Considerations As part of this work, we are releasing pre-trained bilingual models. Big language models can be misused in many different ways including spreading of fake news, generation of toxic speech, etc. We encourage the readers to refer to Bender et al. [36], Brown et al. [37] for an extended discuss of the risks of releasing powerful language models. In our case, the released models are not nearly as big or powerful as state-of-the-art networks such as GPT-3. For all our experiments, we are using the small version of GPT-2 and the main objective is to see how learning multiple languages affects robustness to weight corruptions. Additionally, we are not training these models from scratch, but we are using the recycling technique proposed in de Vries and Nissim [13], hence the environmental cost of the training is much smaller.

F Code and License

We include the code in the Supplementary Material. We plan to open-source the code upon paper acceptance. The code will be released under the GNU GENERAL PUBLIC LICENCE. The interested reader should also refer to the licenses of pre-existing software we use. Please look at the `requirements.txt` file of our code to find all our dependencies.

The code for the training of the bilingual models is written in PyTorch [38] and it is based on the implementation of GPT-2 found in the `transformers` [39] library.

The code for the linear experiments is written in JAX [40]. We plan to include support for PyTorch for the official release.

Apart from the release of the code, we will share publicly all our final pre-trained models. We expect that the release of bilingual and monolingual models trained on identical conditions will motivate further research in this area by cognitive scientists doing computational research. The main motivation for this paper was a theory from Cognitive Science regarding increased Cognitive Reserve in bilingual people. We expect that there could be many more interesting directions in Cognitive Science that can be studied from a computational perspective and we hope that the release of bilingual models will contribute towards this goal.

References

[1] John AE Anderson, John G Grundy, Cheryl L Grady, Fergus IM Craik, and Ellen Bialystok. Bilingualism contributes to reserve and working memory efficiency: Evidence from structural and functional neuroimaging. *Neuropsychologia*, 163:108071, 2021.

[2] Brian T Gold, Nathan F Johnson, and David K Powell. Lifelong bilingualism contributes to cognitive reserve against white matter integrity declines in aging. *Neuropsychologia*, 51(13): 2841–2846, 2013.

[3] Ellen Bialystok, Fergus IM Craik, and Morris Freedman. Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, 45(2):459–464, 2007.

[4] Fergus IM Craik, Ellen Bialystok, and Morris Freedman. Delaying the onset of alzheimer disease: Bilingualism as a form of cognitive reserve. *Neurology*, 75(19):1726–1729, 2010.

[5] Daniel Barulli and Yaakov Stern. Efficiency, capacity, compensation, maintenance, plasticity: emerging concepts in cognitive reserve. *Trends in cognitive sciences*, 17(10):502–509, 2013.

[6] Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1135–1143, Cambridge, MA, USA, 2015. MIT Press.

[7] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17(1):2853–2884, jan 2016. ISSN 1532-4435.

[8] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[9] Gene H Golub and Charles F Van Loan. Matrix computations. edition, 1996.

[10] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[13] Wietse de Vries and Malvina Nissim. As good as new. how to successfully recycle english gpt-2 to make models for other languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. doi: 10.18653/v1/2021.findings-acl.74. URL http://dx.doi.org/10.18653/v1/2021.findings-acl.74.

[14] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

[15] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

[16] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context, 2016.

[17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018. doi: 10.18653/v1/w18-5446. URL http://dx.doi.org/10.18653/v1/w18-5446.

[18] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[19] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017. doi: 10.18653/v1/s17-2001. URL http://dx.doi.org/10.18653/v1/S17-2001.

[20] Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. First quora dataset release: Question pairs, 2017. URL https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs.

[21] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*, 2012.

[22] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

[23] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice, 2006.

[24] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics, 2007.

[25] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009.

[26] William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[27] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[28] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. doi: 10.18653/v1/d16-1264. URL `http://dx.doi.org/10.18653/v1/D16-1264`.

[29] Chengzhi Mao, Amogh Gupta, Vikram Nitin, Baishakhi Ray, Shuran Song, Junfeng Yang, and Carl Vondrick. Multitask learning strengthens adversarial robustness. *Lecture Notes in Computer Science*, page 158–174, 2020. ISSN 1611-3349. doi: 10.1007/978-3-030-58536-5_10. URL `http://dx.doi.org/10.1007/978-3-030-58536-5_10`.

[30] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[31] Salah Ghamizi, Maxime Cordy, Mike Papadakis, and Yves Le Traon. Adversarial robustness in multi-task learning: Promises and illusions, 2021.

[32] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1990.

[33] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[34] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Long live the lottery: The existence of winning tickets in lifelong learning. In *International Conference on Learning Representations*, 2021.

[35] Michel Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

[36] 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*, pages 610–623, 2021.

[37] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy,

Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[39] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2019.

[40] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.