Analysis of Sampling Strategies for Multi-Task Learning in Transformer Models

Anirudh Lakhotia*
PES Center for Pat. Recog.
Dept. of Comp. Sc. and Engg.
PES University, Bengaluru, India
anirudhlakhotia5@gmail.com

Akash Kamalesh*
PES Center for Pat. Recog.
Dept. of Comp. Sc. and Engg.
PES University, Bengaluru, India
akash.kamalesh03@gmail.com

Nischal Helagally Shantharaju*
PES Center for Pat. Recog.
Dept. of Comp. Sc. and Engg.
PES University, Bengaluru, India
nischalhelagally@gmail.com

Prerana Sanjay Kulkarni*
PES Center for Pat. Recog.
Dept. of Comp. Sc. and Engg.
PES University, Bengaluru, India
prer.kulk@gmail.com

Gowri Srinivasa*
PES Center for Pat. Recog.
Dept. of Comp. Sc. and Engg.
PES University, Bengaluru, India
gsrinivasa@pes.edu

***Abstract*—Multitask learning has emerged as a powerful paradigm for enhancing natural language understanding capabilities in large language models. However, the effectiveness of multitask learning heavily depends on the sampling strategy used to balance exposure to tasks of varying sizes and complexities during training. In this work, we conduct a comprehensive empirical analysis of three sampling strategies using T5-small on a subset of the GLUE benchmark: examples-proportional sampling, which samples based on raw dataset sizes; temperature-scaled sampling with T=10.0, which moderates size-based differences; and equal sampling, which gives uniform probability to all tasks. Our analysis examines the impact of these strategies on model performance, convergence patterns, task representation, and internal model dynamics through spectral analysis. Our results demonstrate that temperature-scaled sampling provides a strong balance between task representation and overall performance, while equal sampling achieves the highest average score across all tasks. Spectral analysis reveals that sampling strategies produce fundamentally different internal representations despite similar performance metrics, with higher temperatures promoting more uniform and stable layer behavior. We provide practical guidelines for selecting appropriate sampling strategies based on dataset characteristics, computational constraints, and specific training goals, contributing to more effective multitask learning approaches for language models.**

***Index Terms*—multitask learning, sampling strategies, language models, T5, GLUE benchmark, natural language processing**

I. Introduction

Recent advances in large language models (LLMs) have highlighted the critical importance of generalization capabilities, particularly in scenarios outside the training distribution. Models like Deepseek-R1 have achieved remarkable scores on generalization benchmarks [1], demonstrating the ability to tackle novel scenarios with minimal or no examples. This capability becomes increasingly crucial as we progress toward more advanced AI systems that must adapt to new tasks by leveraging knowledge from related experiences, similar to human learning patterns. Multitask learning (MTL) has emerged as a powerful paradigm for enhancing such generalization capabilities, where models are trained simultaneously on multiple tasks to capture underlying patterns and shared representations rather than overfitting to task-specific features [2], [3].

The effectiveness of multitask learning, however, heavily depends on how data from different tasks is sampled or weighted during training. This strategy becomes particularly critical when dealing with datasets of varying sizes and complexities, as inappropriate strategies can lead to catastrophic forgetting or poor performance on smaller tasks [4], [5], and recent work explores advanced weighting/mixing strategies [16], [17], indicating that the choice of strategy can significantly impact both the model's final performance and its learning dynamics [6].

In this work, we conduct a comprehensive analysis of three fundamental sampling strategies using T5-small on the GLUE benchmark [7]: examples-proportional sampling, which samples based on raw dataset sizes; temperature-scaled sampling with T=10.0, which moderates size-based differences; and equal sampling, which gives uniform probability to all tasks regardless of size. We selected these three sampling strategies due to their widespread use and demonstrated effectiveness in previous multitask learning studies [4], [5], [8]. Our investigation examines the complex relationships between sampling strategies, task characteristics, and model performance.

Our contributions include:

- A systematic empirical comparison of three sampling strategies in multitask learning, including detailed analysis of their effects on both high-resource and low-resource tasks.
- Novel insights into the relationship between sampling strategies and model convergence, supported by extensive experimental evidence.

[0*]These authors contributed equally to this work.

- Analysis of how temperature-scaled sampling affects the learning dynamics and internal representations of multitask models through spectral analysis.
- Practical guidelines for selecting appropriate sampling strategies based on task characteristics, dataset sizes, and computational constraints.

II. Related Work

A. Multitask Learning in NLP

Multitask learning has a rich history in natural language processing, dating back to early work on shared representations [2]. The core principle involves training models on multiple related tasks simultaneously, allowing them to leverage commonalities and transfer knowledge across tasks. This approach has proven particularly effective in NLP, where different tasks often share underlying linguistic knowledge and representations. For recent comprehensive surveys covering traditional, deep, and foundation model approaches, see [14], [15].

Recent advances in transformer-based architectures have renewed interest in multitask learning, with models like T5 [4] and MASS [9] demonstrating impressive performance across diverse NLP tasks. These models have introduced several key innovations in handling multiple tasks. Task-specific adapters [10] allow efficient parameter sharing while maintaining task-specific specialization. Prompt-based methods [11] have emerged as a powerful approach for framing all tasks in a consistent format. Additionally, sophisticated mixing strategies [6] have been developed for balancing different tasks during training. The success of these approaches has highlighted the importance of proper task mixing and sampling strategies in multitask learning scenarios.

B. Sampling Strategies and Training Dynamics

The challenge of balancing multiple tasks during training has led to the development of various sampling approaches. Examples-proportional mixing, initially proposed in [4], samples tasks according to their dataset sizes. While intuitive, this approach can lead to underrepresentation of smaller tasks, potentially compromising their performance. Temperature-scaled mixing [5] addresses this limitation by introducing a temperature parameter to adjust the sampling distribution. This provides more control over task exposure and has become increasingly popular in recent work. The temperature parameter allows practitioners to smoothly interpolate between proportional and uniform sampling, offering a flexible way to handle dataset size disparities.

Equal mixing strategies [8] represent another approach, giving equal importance to all tasks regardless of dataset size. While this ensures adequate representation of smaller tasks, it can lead to inefficient use of larger datasets and slower convergence on high-resource tasks. Recent studies like Hyperformer [12] have shown that temperature values around T=10.0 provide a good balance between proportional and equal sampling for GLUE tasks. While fixed temperature or proportional sampling are common, recent research explores more dynamic approaches, such as uncertainty-aware task weighting [16] or adaptive mixing based on gradient information [17].

III. Methodology

A. Experimental Setup

We conduct our experiments using T5-small [4], a compact encoder-decoder transformer model with approximately 60 million parameters. This model size allows us to run multiple experimental configurations while maintaining reasonable computational requirements. All experiments are conducted using PyTorch 2.1.0 on NVIDIA A100 GPUs with 40GB of memory. We use the Hugging Face Transformers library (version 4.18.0) for model implementation and training infrastructure.

For experiment tracking and analysis, we employ Weights & Biases (wandb) to log training metrics, hyperparameters, and resource utilization. This enables systematic comparison across different sampling strategies and facilitates reproducibility. We also integrate WeightWatcher [13] to extract diagnostic metrics about the model's internal representations during training, providing insights into generalization capabilities and potential overfitting. [1]

B. Dataset Preparation

We use the General Language Understanding Evaluation (GLUE) benchmark [7] as our experimental testbed. GLUE consists of nine diverse natural language understanding tasks, including single-sentence classification, similarity and paraphrase detection, and natural language inference. The tasks vary significantly in size, from approximately 2,500 examples (RTE) to over 390,000 examples (MNLI), making it an ideal benchmark for studying sampling strategies in imbalanced multitask settings.

To ensure consistent preprocessing across tasks, we follow the T5 approach of casting all tasks into a text-to-text format. For each task, we define a specific prompt template that frames the task as a text generation problem. For example, for the SST-2 sentiment classification task, we use the prompt "sentiment: [input]" and train the model to generate "positive" or "negative" as appropriate. This unified approach allows us to handle all tasks within the same model architecture without task-specific modifications.

For computational efficiency, we use a stratified 50% subset of each GLUE task while maintaining the original class distributions. This reduction allows us to run more experimental configurations while still preserving the essential characteristics of each dataset. We verify through preliminary experiments that the performance trends observed on these subsets are consistent with those on the full datasets.

For evaluation, we use task-specific metrics as defined in the GLUE benchmark: accuracy for SST-2, MNLI, QNLI, and RTE; F1 score and accuracy for QQP and MRPC;

[1] All code, scripts, and plots required to reproduce our experiments and analyses are publicly available at: https://github.com/anirudhlakhotia/multitask-sampling-strategies

Matthews correlation coefficient (MCC) for CoLA; and Pearson/Spearman correlation for STS-B. We report both task-specific metrics and an average score across all tasks to assess overall performance.

C. Sampling Strategies

We implement and evaluate three fundamental sampling strategies for multitask learning:

1) Examples-Proportional Sampling: In examples-proportional sampling, the probability of selecting a task is directly proportional to its dataset size. For a task i with dataset size D_i, the sampling probability is:

$$p(i) = \frac{D_i}{\sum_j D_j} \tag{1}$$

This approach naturally allocates more training iterations to larger tasks, which can be beneficial when larger datasets contain more diverse information. However, it may lead to underrepresentation of smaller tasks, potentially compromising their performance.

2) Temperature-Scaled Sampling: Temperature-scaled sampling introduces a temperature parameter T to control the balance between dataset sizes. The sampling probability for task i is:

$$p(i) = \frac{(D_i/D_{\text{total}})^{1/T}}{\sum_j (D_j/D_{\text{total}})^{1/T}} \tag{2}$$

where D_{total} is the sum of all dataset sizes. This formulation has several important properties:

- When $T = 1$, it reduces to examples-proportional sampling.
- As $T \to \infty$, it approaches equal sampling.
- As $T \to 0$, it increasingly favors larger datasets.

The temperature parameter provides a flexible way to control task exposure during training, allowing practitioners to find an optimal balance between large and small tasks. In our experiments, we use a temperature value of $T = 10.0$, which has been used extensively in prior work on multitask learning with GLUE tasks [12].

3) Equal Sampling: Equal sampling assigns the same probability to all tasks regardless of their dataset sizes:

$$p(i) = \frac{1}{N} \tag{3}$$

where N is the total number of tasks. This approach ensures that all tasks receive equal representation during training, which can be particularly important for smaller tasks that might otherwise be overwhelmed. However, it may lead to inefficient use of larger datasets and potentially cause overfitting on smaller tasks due to repeated exposure to the same examples.

In practice, we implement equal sampling by setting the temperature parameter to a very large value ($T = 10^{11}$), effectively approximating infinity. This ensures a uniform sampling distribution across all tasks.

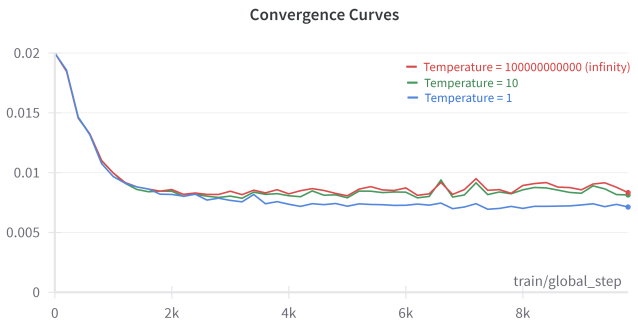

Fig. 1. Initial convergence curves of different sampling strategies for GLUE. This is obtained by running evaluation every 100 steps for the first 10,000 steps on a random subset of the validation set.

IV. Experiments

A. Performance Comparison

We evaluate the performance of each sampling strategy on individual GLUE tasks as well as the average performance across all tasks. For each strategy, we report task-specific metrics as defined in the GLUE benchmark: accuracy for SST-2, MNLI, QNLI, and RTE; F1 score and accuracy for QQP and MRPC; Matthews correlation coefficient (MCC) for CoLA; and Pearson/Spearman correlation for STS-B. Table I presents a comparison of these results.

Table I shows that examples-proportional sampling excels on high-resource tasks, while equal sampling achieves the highest average performance across tasks.

B. Convergence Patterns

To understand how different sampling strategies affect training dynamics, we track the validation performance during the training process. Fig. 1 illustrates the validation performance on GLUE tasks during the initial stages of training for the three sampling strategies. Examples-proportional sampling converges faster initially, whereas equal sampling converges slower but achieves better long-term performance on smaller tasks.

C. Task Exposure Distribution

We track the proportion of examples from each task seen during training to quantify how different sampling strategies allocate training resources. Table II shows the actual task exposure rates for each sampling strategy compared to the original dataset distribution.

D. Spectral Analysis Methodology

To analyze internal model representations, we employ WeightWatcher, a neural network analysis tool that examines the spectral properties of weight matrices across different layers and training configurations [13]. WeightWatcher computes various metrics derived from the eigenvalue and singular value distributions of each layer's weight matrix, providing insights into model generalization capabilities, learning dynamics, and internal representations.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT SAMPLING STRATEGIES ON GLUE TASKS (USING 50% SUBSETS)

Sampling Strategy	**CoLA**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STS-B**	**Avg**
Examples-Proportional (T=1)	0.1041	**0.8013**	**0.8529/0.8529**	**0.8989**	**0.8742/0.8825**	0.6522	**0.9083**	0.8452/0.8467	0.7572
Temperature-Scaled (T=10.0)	0.1042	0.7931	0.8235/0.8235	0.8946	0.8613/0.8709	**0.7246**	0.8945	**0.8555/0.8550**	0.7807
Equal (T=10^{11})	**0.1767**	0.7902	0.8333/0.8333	0.8931	0.8611/0.8707	**0.7246**	0.8945	0.8541/0.8575	**0.8215**

TABLE II
TASK EXPOSURE DISTRIBUTION ACROSS SAMPLING STRATEGIES (BASED ON 50% SUBSETS)

Sampling Strategy	**CoLA**	**MNLI**	**MRPC**	**QNLI**	**QQP**	**RTE**	**SST-2**	**STS-B**
Original Dataset (%)	0.90	41.51	0.39	10.97	38.35	0.26	7.01	0.61
Examples-Proportional (T=1)	0.90	41.51	0.39	10.97	38.35	0.26	7.01	0.61
Temperature-Scaled (T=10)	10.87	15.94	9.99	13.96	15.82	9.61	13.35	10.45
Equal (T=10^{11})	12.50	12.50	12.50	12.50	12.50	12.50	12.50	12.50

We track the following key metrics throughout training:

- Alpha Power Law: Measures the power-law decay of singular values, indicating how hierarchical or structured the learned features are.
- Maximum Eigenvalue: The largest eigenvalue of the weight matrix, indicating the maximum amplification applied to inputs.
- Spectral Norm: A measure of the layer's sensitivity to input perturbations.
- Effective Rank: Estimates the number of significant singular values, indicating the effective dimensionality of representations.

We collect these metrics at regular intervals during training for all layers across the three sampling strategies. This approach allows us to analyze how different temperature settings affect the model's internal dynamics and representation formation beyond what is apparent from task performance metrics alone.

V. ANALYSIS

A. Performance Trade-offs Across Sampling Strategies

Table I clearly demonstrates distinct performance trade-offs among the sampling strategies. Examples-proportional sampling achieves the highest performance on high-resource tasks such as MNLI (0.8013), QQP (0.8742/0.8825), QNLI (0.8989), MRPC (0.8529), and SST-2 (0.9083). This aligns with expectations, as this strategy allocates most training iterations to larger datasets, enabling rapid specialization.

Temperature-scaled sampling (T=10.0) provides a balanced performance profile, achieving competitive results across both high- and low-resource tasks. It notably excels on STS-B (0.8555/0.8550) and shares the top performance on RTE (0.7246). This balanced allocation of training resources moderates the dominance of larger tasks while still leveraging their diversity.

Equal sampling achieves the highest overall average score (0.8215), despite not dominating most individual tasks. It significantly improves performance on smaller tasks such as CoLA (0.1767) and RTE (0.7246). This suggests that equal representation across tasks fosters more robust general-purpose representations, beneficial for scenarios requiring balanced performance across diverse tasks.

B. Learning Dynamics and Convergence Behavior

Figure 1 illustrates distinct convergence patterns across sampling strategies. Examples-proportional sampling demonstrates the fastest initial convergence, benefiting from frequent exposure to larger datasets that provide diverse learning signals early in training. However, this strategy plateaus earlier on smaller tasks due to limited exposure.

Temperature-scaled sampling (T=10.0) exhibits moderate convergence speed, balancing rapid improvement on large tasks with steady progress on smaller ones. This balanced approach develops robust representations across the task spectrum over time.

Equal sampling shows the slowest initial convergence but achieves the strongest long-term performance on smaller tasks. This delayed convergence is expected, as the model spends proportionally more time on smaller datasets, which initially provide fewer diverse learning signals but ultimately lead to better generalization.

C. Impact of Task Exposure Distribution

Table II highlights how sampling strategies fundamentally reshape task exposure during training. Examples-proportional sampling mirrors the highly skewed original dataset distribution, allocating over 79% of training exposure to MNLI and QQP alone, while smaller tasks collectively receive less than 3%. This explains the strategy's strong performance on high-resource tasks and weaker performance on low-resource tasks.

Temperature-scaled sampling (T=10.0) significantly moderates this imbalance, allocating between 9.61% (RTE) and 15.94% (MNLI) exposure across tasks. This balanced distribution explains its competitive performance across both large and small tasks.

Equal sampling uniformly allocates 12.5% exposure to each task, dramatically increasing exposure to smaller tasks (e.g., RTE exposure increases from 0.26% to 12.5%). This explains the substantial performance improvements on smaller tasks

TABLE III
SPECTRAL PROPERTIES AT FINAL TRAINING STEP (50,000 STEPS)

Metric	T=1	T=10	T=10^{11}	Change*
Mean Alpha	4.41	4.35	4.33	-1.8%
Alpha CV	0.427	0.351	0.343	-19.7%
Max Eigenvalue	36.43	32.18	29.73	-18.4%
Effective Rank	12.38	14.25	15.87	+28.2%
Spectral Norm	5.92	5.47	5.21	-12.0%

*Change calculated from T=1 to T=10^{11}

and the slight performance decrease on larger tasks due to reduced exposure.

D. Spectral Analysis of Internal Representations

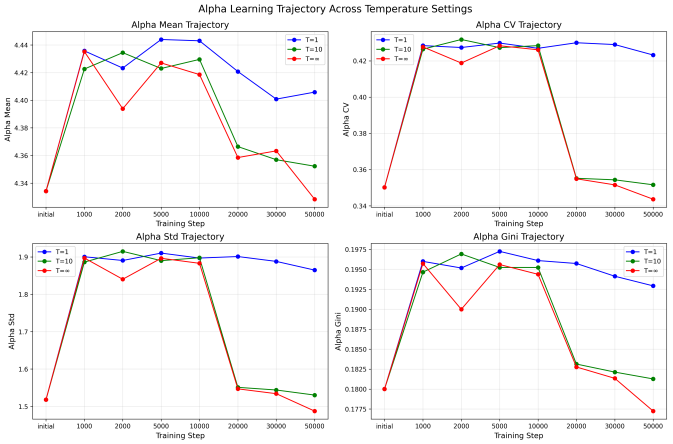

Fig. 2. Evolution of alpha metrics across training steps for different temperature settings. The plots show (clockwise from top-left): mean alpha values, alpha coefficient of variation, alpha Gini coefficient, and alpha standard deviation.

WeightWatcher analysis of the models' weight matrices reveals significant differences in their internal organization across temperature settings. These spectral metrics (Table III, Figure 2) provide insights into the models' generalization capabilities and representational efficiency.

1) Power Law Analysis: The alpha parameter, which measures the power-law decay of singular values in weight matrices, shows systematic changes with temperature (Figure 2):

- The mean alpha decreases slightly from 4.41 (T=1) to 4.33 (T=10^{11}), indicating a potential shift toward more distributed representations at higher temperatures.
- Alpha CV (Coefficient of Variation) drops significantly by 19.7% from T=1 to T=10^{11}, suggesting much more uniform behavior across layers at higher temperatures.
- This trend toward more uniform, distributed representations typically correlates with better generalization.

2) Stability and Capacity Metrics: The spectral properties reveal important trends in model stability and capacity utilization:

- **Maximum Eigenvalue** decreases by 18.4% as temperature increases, indicating reduced sensitivity to input perturbations. Lower maximum eigenvalues suggest better numerical stability and robustness to input variations, a key factor in generalization performance.
- **Spectral Norm** shows a 12.0% reduction, further confirming improved stability at higher temperatures. Lower spectral norms are associated with better Lipschitz constraints and more robust generalization bounds.
- **Effective Rank** increases substantially (+28.2%) with temperature, suggesting that higher temperatures lead to more distributed representations using a larger portion of the model's capacity. Higher effective rank typically correlates with better generalization as it indicates the model is learning more diverse and comprehensive features.

3) Implications for Generalization: These spectral characteristics suggest that higher temperatures promote several properties associated with better generalization:

- **Improved Stability**: The combination of lower maximum eigenvalues and spectral norms indicates more stable and robust representations, making the model less sensitive to input perturbations.
- **Better Capacity Utilization**: The increased effective rank suggests more efficient use of model capacity, with representations distributed across more dimensions rather than concentrated in a few dominant directions.
- **Balanced Layer Behavior**: The reduced alpha CV indicates more uniform behavior across layers, potentially preventing over-specialization and promoting more robust feature hierarchies.

These findings align with the observed performance patterns, where higher temperatures lead to more balanced performance across tasks and potentially better generalization to unseen tasks. The spectral metrics suggest that temperature-based sampling acts as an implicit regularizer, promoting more stable and distributed representations while maintaining task performance.

E. Practical Implications and Recommendations

Our analysis provides clear practical guidelines for selecting appropriate sampling strategies based on specific training goals and constraints:

- **Examples-Proportional Sampling (T=1)**: Recommended when rapid convergence and high performance on large, high-resource tasks are priorities. Ideal for scenarios with limited computational resources or when smaller tasks are less critical.
- **Temperature-Scaled Sampling (T=10.0)**: Recommended for balanced performance across diverse tasks. This strategy effectively moderates task exposure, achieving competitive results on both large and small tasks, making it suitable for general-purpose multitask learning scenarios.
- **Equal Sampling (T=10^{11})**: Recommended when robust generalization and balanced performance across all tasks are critical, despite potentially slower initial convergence. Ideal for scenarios where smaller tasks are equally important or when generalization to new tasks is a priority.

Overall, our analysis demonstrates that sampling strategies significantly impact both task-specific performance and internal model dynamics. Practitioners should carefully select sampling strategies based on their specific priorities, dataset characteristics, and computational constraints.

VI. Limitations and Future Work

While our study provides valuable insights into multitask sampling strategies, there are a few limitations worth noting:

- Our experiments focus exclusively on the GLUE benchmark using the T5-small model; future work could explore generalization to larger models and more diverse task sets.
- We employ fixed temperature/sampling; future work could investigate adaptive strategies [16], [17].

Addressing these limitations by exploring larger models, diverse task sets, and adaptive sampling strategies represents a promising direction for future research in multitask learning.

VII. Conclusion

Our comprehensive analysis of multitask sampling strategies reveals distinct trade-offs between examples-proportional, temperature-scaled, and equal sampling approaches. Each strategy offers unique advantages: examples-proportional sampling excels on high-resource tasks and demonstrates faster initial convergence; temperature-scaled sampling (T=10.0) achieves strong performance on low-resource tasks while maintaining competitive results on larger ones; and equal sampling yields the highest overall average score (0.8215) despite slower initial convergence.

Beyond performance metrics, our spectral analysis uncovered a critical insight: these sampling strategies produce fundamentally different internal representations. Equal sampling leads to more distributed representations with higher effective rank, while examples-proportional sampling creates more specialized embeddings with higher maximum eigenvalues. Temperature acts as an implicit regularizer, with higher values promoting more uniform and stable layer behavior.

Based on these findings, we recommend selecting sampling strategies according to specific requirements: examples-proportional sampling for rapid prototyping or when high-resource task performance is critical; temperature-scaled sampling for balanced performance across diverse tasks; and equal sampling for maximizing generalization capabilities despite potentially longer training times. These insights contribute to more effective multitask learning approaches for language models and highlight the importance of appropriate sampling strategies in multi-task scenarios.

Acknowledgment

We would like to thank PES University for providing the computational resources and infrastructure necessary for conducting this research.

References

[1] D. Guo, J. Huang, M. Qi, Z. Lin, S. Zhang, P. Chen, H. Cheng, Z. Dai, R. Zheng, Y. Song, et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[2] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, pp. 41–75, 1997.

[3] Z. Zhang, W. Yu, M. Yu, Z. Guo, and M. Jiang, "A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods," in *Proc. 17th Conf. European Chapter Assoc. Comput. Linguistics (EACL)*, May 2023, pp. 943–956. [Online]. Available: https://aclanthology.org/2023.eacl-main.66/

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[5] J. Wang, K. Toutanova, J. H. Lee, and G. W. Cherry, "Gradient-based Analysis of NLP Models is Manipulable," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Nov. 2020, pp. 3402–3410. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.305

[6] A. Aghajanyan, A. Gupta, G. Singh, V. Stoyanov, S. Goyal, M. Lewis, L. Zettlemoyer, and S. Peshterliev, "Muppet: Massive Multi-task Representations with Pre-Finetuning," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2021, pp. 10197–10215. [Online]. Available: https://aclanthology.org/2021.emnlp-main.798/

[7] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=rJ4km2R5t7

[8] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1871–1880.

[9] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MASS: Masked Sequence to Sequence Pre-training for Language Generation," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 5926–5936.

[10] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *International Conference on Machine Learning (ICML)*. PMLR, 2019, pp. 2790–2799.

[11] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=gEZrGCozdq

[12] R. Karimi Mahabadi, J. Henderson, and S. Ruder, "Compacter: Efficient low-rank hypercomplex adapter layers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 1022–1035.

[13] C. H. Martin, T. Peng, and M. W. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Communications*, vol. 12, no. 1, p. 4122, 2021.

[14] Z. Zhang, W. Yu, M. Yu, Z. Guo, and M. Jiang, "A Survey of Multi-task Learning in Natural Language Processing: Regarding Task Relatedness and Training Methods," in *Proc. 17th Conf. European Chapter Assoc. Comput. Linguistics (EACL)*, May 2023, pp. 943–956. [Online]. Available: https://aclanthology.org/2023.eacl-main.66/

[15] J. Yu, Y. Dai, X. Liu, J. Huang, Y. Shen, K. Zhang, Y. Chen, and Y. Sun, "Unleashing the Power of Multi-Task Learning: A Comprehensive Survey Spanning Traditional, Deep, and Pretrained Foundation Model Eras," *arXiv preprint arXiv:2402.01113*, 2024.

[16] Z. Shi, Z. Li, C. Wu, Z. Wang, Y.-C. Wang, and B. Wang, "Uncertainty-Aware Dynamic Task Weighting for Multi-Task Natural Language Understanding," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096865.

[17] C. Xu, J. Zhou, M. Yu, A. Aghajanyan, N. Goyal, A. Metallinou, H. He, and A. Srinivasan, "PiKE: Adaptive Data Mixing for Multi-Task Learning Under Low Gradient Conflicts," *arXiv preprint arXiv:2502.06244*, 2025.