A APPENDIX

A.1 JUSTIFYING THE INTRODUCTION OF A META-HEAD

Proof. To arrive at Equation 7 we start with the closed form solution for $\nabla_{w_i}\mathcal{L}^{val}_{T^*}(\theta^*)$ and then introduce approximations in order to produce Equation 7. First, note that :

$$\frac{\partial \mathcal{L}^{val}_{T^*}(\theta^*(\mathbf{w}))}{\partial w_i} = \Big(\nabla_\theta \mathcal{L}^{val}_{T^*}(\theta^*(\mathbf{w}))\Big)^T \Big(\nabla_{w_i}\theta^*(\mathbf{w})\Big) \quad \texttt{[Chain rule]} \tag{8}$$

To get $\nabla_{w_i}\theta^*(\mathbf{w})$ we invoke the Cauchy Implicit Function Theorem (IFT) as with Lorraine et al. (2020); Navon et al. (2020); Liao et al. (2018):

$$\begin{aligned}
\nabla_{w_i}\theta^*(\mathbf{w}) &= \Big[\nabla^2_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big]^{-1}\Big[\nabla_{w_i}\nabla_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big] \quad \texttt{[IFT]} \\
&= \Big[\nabla^2_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big]^{-1}\Big[\nabla_{w_i}\nabla_\theta \Big(w^*\mathcal{L}_{T^*}(\theta^*(\mathbf{w})) + \sum_{T_i \in \mathbb{T}_{\text{aux}}} w_i \mathcal{L}_{T_i}(\theta^*(\mathbf{w}))\Big)\Big] \\
&= \Big[\nabla^2_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big]^{-1}\Big[\nabla_\theta \mathcal{L}_{T_i}(\theta^*(\mathbf{w}))\Big] \quad \texttt{[Only terms with } w_i \texttt{ survive]}
\end{aligned}$$

Bringing it all together, we get :

$$\frac{\partial \mathcal{L}^{val}_{T^*}(\theta^*(\mathbf{w}))}{\partial w_i} = \Big(\nabla_\theta \mathcal{L}^{val}_{T^*}(\theta^*(\mathbf{w}))\Big)^T \Big(\Big[\nabla^2_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big]^{-1}\Big[\nabla_\theta \mathcal{L}_{T_i}(\theta^*(\mathbf{w}))\Big]\Big) \tag{9}$$

□

Computing $\nabla_{w_i}\mathcal{L}^{val}_{T^*}(\theta^*)$ from Equation 9 is computationally unwieldy since we would not only have to optimize θ to convergence for every step of w_i but we would also have to invert the Hessian of a typically large model. Our middle ground between Equations 9 and 6 (Equation 7) makes use of the following approximations:

- We approximate the inverse Hessian with the identity. This approximation is not new; we follow previous work like Lorraine et al. (2020)(Table 3) who explore the use of this approximation because of computational efficiency.

  $$\Big[\nabla^2_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big]^{-1} = \lim_{i\to\infty}\sum_{j=0}^{i}\Big(\mathbf{I} - \nabla^2_\theta \mathcal{L}_{\text{total}}(\theta^*(\mathbf{w}))\Big)^j \approx \mathbf{I}$$

  We are assuming the contribution of terms with $i > 0$ are negligible.

- Instead of training the whole network to convergence, at each time-step, we fix the body of the network and train a special head ϕ^* to convergence on a small batch of end-task training data. We then use $[\theta_{\text{body}}; \phi^*]$ as a proxy for θ^*. This is a computationally feasible work-around to training all of θ to convergence to get a single step gradient estimate. Especially in the continued pre-training setting where a pre-trained *generalist* model like BERT is used as θ_{body}, this approximation is reasonable. To our knowledge, we are the first to suggest this approximation.

  $$\nabla_\theta \mathcal{L}^{val}_{T^*}(\theta^*) \rightarrow \nabla_\theta \mathcal{L}^{val}_{T^*}([\theta_{\text{body}}; \phi^*])$$

- Above, we have approximated $\theta^* = [\theta_{\text{body}}; \phi^*]$. Since ϕ^* is only used to evaluate end-task (T^*) validation data, it means θ remains unchanged with respect to the training data for task T_i. Thus $\nabla_\theta \mathcal{L}_{T_i}([\theta_{\text{body}}; (\phi^*, \ldots, \phi^i)]) = \nabla_\theta \mathcal{L}_{T_i}([\theta_{\text{body}}; \phi^i]) = \nabla_\theta \mathcal{L}_{T_i}(\theta)$

Bringing it all together, we get Equation 7, repeated here:

$$\frac{\partial \mathcal{L}^{val}_{T^*}(\theta^*(\mathbf{w}))}{\partial w_i} \approx \big(\nabla_\theta \mathcal{L}_{T_i}\big)^T \big(\nabla_\theta \mathcal{L}^{val}_{T^*}([\theta_{\text{body}}; \phi^*]_t)\big)$$

A.2 CALCULATING P-VALUES FROM PERMUTATION TEST

We used the permutation test (Good, 2005; Dror et al., 2018) to test for statistical significance. For each test, we generate 10000 permutations to calculate significance level. This is sufficient to converge to a stable p-value without being a computational burden. We chose this over the common student t-test because :

1. We have only 10 runs per algorithm and permutation tests are more robust at low sample size
2. Permutation test is assumption free. Student t-tests assume that the samples are normally distributed
3. Permutation test is robust to variance in the samples, so even though error-bars can overlap, we still establish significant differences in the samples. Variance in our results is expected due to small dataset sizes of end-tasks.

A.3 ALGORITHM FOR META-TARTAN

Algorithm 1: End-task Aware Training via Meta-learning (META-TARTAN)

Require: T^*, $\mathbf{T}_{\text{aux}}$: End-task, Set of auxiliary pre-training tasks
Require: η, β_1, β_2: Step size hyper-parameters
Initialize :
  Pre-trained RoBERTa as shared network body, θ_{body}
  Task weightings: $w^*, w_i = \frac{1}{|\mathbf{T}_{\text{aux}}|+1}$
Randomly initialize :
  end-task head as ϕ'
  meta head for end-task as ϕ^*
  task head, ϕ^i, for each $T_i \in \mathbf{T}_{\text{aux}}$
while *not done* **do**
  $B^*_{\text{tr}} \sim T^*_{\text{train}}$ `// Sample a batch from end-task`
  $g^*_\theta, g^*_\phi \leftarrow \left[\nabla_\theta, \nabla_{\phi'}\right]\left(\mathcal{L}_{T^*}(\theta, \phi', B^*_{\text{tr}})\right)$ `// Get end-task grads`
  $g^i_\theta, g^i_\phi \leftarrow \left[\nabla_\theta, \nabla_{\phi^i}\right]\left(\mathcal{L}_{T_i}(\theta, \phi^i, B_i)\right)$ `// Get task grads.` $\forall i \in [n],\ B_i \sim T_i$
  `// Learn a new meta head`
  $\phi^* \leftarrow$ `estimate_meta_head`$(B^*_{\text{tr}}, \beta_2, \theta, \phi^*)$ `//` $B^*_{\text{tr}} \sim T^*_{\text{train}}$
  $g^*_{meta} \leftarrow \nabla_\theta \mathcal{L}_{T^*}(\theta, \phi^*, B^*_{\text{val}})$ `//` $B^*_{\text{val}} \sim T^*_{\text{val}}$
  `// Update task weightings`
  $w^* \leftarrow w^* + \eta \cos(g^*_{meta}, g^*_\theta)$
  $w_i \leftarrow w_i + \eta \cos(g^*_{meta}, g^i_\theta)$
  `// Update task parameters`
  $\alpha^*, \alpha_1, \ldots, \alpha_{|\mathbf{T}_{\text{aux}}|} = \text{softmax}(w^*, w_1, \ldots, w_{|\mathbf{T}_{\text{aux}}|})$
  Update $\theta_{\text{body}} \leftarrow \theta_{\text{body}} - \beta_1\left(\alpha^* g^*_\theta + \sum_i \alpha_i g^i_\theta\right)$
  Update $\left(\phi_i \leftarrow \phi_i - \beta_2 g^i_\phi\right), \left(\phi' \leftarrow \phi' - \beta_2 g^*_\phi\right)$
end
Result : θ, ϕ'

A.4 VISION EXPERIMENTS

We validate that the gains from end-task Aware Training are not siloed to only learning from text. We conduct an experiment comparing end-task aware training on images to its end-task agnostic variant. We use the Cifar100 dataset (Krizhevsky et al., 2009). We use the Medium-Sized Mammals superclass (one of the 20 coarse labels) as our main task whilst the other 19 super classes are used as auxiliary data. Our primary task is thus a 5-way classification task of images different types of

Method	Medium-Sized Mammals
Regular (Task-Agnostic) Pre-training	$46.7_{2.2}$
MT-TARTAN	$51.3_{1.2}$
META-TARTAN	$\mathbf{52.3_{3.8}}$

Table 5: We report averages across 3 random seeds. Best average task accuracy is bolded.

medium-sized mammals whilst whilst the remaining 95 classes are grouped into a single auxiliary task.

As can be seen from Table 5, being end-task aware improves over task agnostic pre-training. We find that, again, when our auxiliary task consist of solely domain data and no task data, META-TARTAN performs better than MT-TARTAN (as measured by averaged performance).

A.5 Full TAPT Table with Significance Levels

We repeat Table 1 and provide details about levels of statistical signifance.

Task	**TAPT**	**MT-TARTAN**	$p-$values	**META-TARTAN**	$p-$values
ACL-ARC	$67.74_{3.68}$	$\mathbf{70.48}_{4.42}$	0.040	$70.08_{4.70}$	0.069
SCIERC	$79.53_{1.93}$	$\mathbf{80.81}_{0.74}$	0.038	$\mathbf{81.48}_{0.82}$	0.005
CHEMPROT	$82.17_{0.065}$	$\mathbf{84.29}_{0.63}$	0.000	$\mathbf{84.49}_{0.50}$	0.000

Table 6: Duplicate of Table 1. Significance levels as computed from the permutation test. All $p-$values are relative to the TAPT column. Statistically significant performance(p-value from permutation test < 0.05), is boldfaced

A.6 Full DAPT/DAPT+TAPT Table

We repeat Table 3 and provide details about levels of statistical signifance.

Task	**DAPT**	**DAPT+TAPT**	**MT-TARTAN**	p-values	**META-TARTAN**	p-values
ACL-ARC	$68.60_{2.62}$	$69.12_{5.76}$	$71.58_{1.65}$	0.110	$71.05_{2.37}$	0.174
SCIERC	$76.44_{1.19}$	$77.62_{1.38}$	$\mathbf{81.02_{1.24}}$	0.000	$\mathbf{81.41_{1.70}}$	0.000
CHEMPROT	$80.76_{0.54}$	$78.22_{0.74}$	$\mathbf{83.77_{0.60}}$	0.000	$\mathbf{83.38_{0.89}}$	0.000

Table 7: Duplicate of Table 2. Significance levels as computed from the permutation test. All $p-$values are relative to $\max(\text{DAPT}, \text{DAPT} + \text{TAPT})$. Statistically significant performance(p-value from permutation test < 0.05), is boldfaced